Signal Classification Techniques for Searches and Measurements at the LHC

PETROV, PETAR,MARINOV

Signal Classification Techniques for Searches and Measurements at the LHC

Petar Marinov Petrov

A Thesis presented for the degree of
Doctor of Philosophy

Institute of Particle Physics Phenomenology
Department of Physics
University of Durham
England

September 2016

Dedicated to

baba

Signal Classification Techniques for Searches and Measurements at the LHC

Petar Marinov Petrov

Submitted for the degree of Doctor of Philosophy
September 2016

Abstract

This thesis focuses on three different examples of techniques designed to extract signal from background in the highly polluted by QCD environment of the proton-proton collisions at the Large Hadron Collider. The first is an attempt at quark-gluon tagging with the help of a simplified version of the shower deconstruction approximation to the likelihood ratio. We find that it outperforms some frontrunners in the field for a large variety of jet definitions and constraints, assuming topocluster-like objects instead of hadrons as seeds. The second search is tasked with identifying boosted W bosons, emitted from high virtuality quarks, thereby measuring the effects of Sudakov logarithmic enhancement under different assumptions of the systematic uncertainty. Finally we examine the LHC's capability to measure and constrain the strength of the $t\bar{t}H(b\bar{b})$ channel in an extensive search of various modestly boosted phase space regions. Under optimistic assumptions about the missing energy reconstruction in b-tagged jets and the handling of the systematic uncertainty, we are able to exclude deviations on the order of 20% from the SM expectation.

Declaration

The work in this thesis is based on research carried out at the Institute of Particle Physics Phenomenology, Department of Physics, Durham University. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text. Chapter 2 is based on [1] in collaboration with Michael Spannowsky, Davison Soper and Danilo Ferreira de Lima. Chapter 3 is based on research carried out in collaboration with Michael Spannowsky, Frank Krauss and Marek Schonherr and presented in [2]. Finally, Chapter 4 follows the phenomenological part of [3], which was written in collaboration with Michael Spannowsky, Stefano Pozzorini and Niccolo Moretti.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Standard Model (SM) of particle physics describes the strong nuclear, weak nuclear and electromagnetic interactions of the fundamental matter fields (six quarks and six leptons) by treating them as multiplets in a representation of a local gauge symmetry group $SU(3)_c \otimes SU(2)_I \otimes U(1)_Y$ [6–9]. The symmetry requirement necessitates interactions through gauge vector bosons, one for each conserved current. Within this framework, the gauge bosons cannot be massive. Moreover, the weak nuclear force is chiral, it affects the left and right handed components of the fermionic fields differently. Therefore, a mass term $m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ will violate the gauge symmetry. Both of these conditions are in contrast with observations, which is solved by the Higgs Mechanism [10–12]. It introduces a scalar field that transforms in the fundamental representation of the weak isospin group and also has a weak hypercharge. The potential of the Higgs field has a vacuum that breaks the underlying gauge symmetry $SU(2)_I \otimes U(1)_Y \rightarrow U(1)_{\mathrm{EM}}$, giving masses to the $W^{\pm}$ and Z bosons, leaves the photon massless and introduces a massive scalar particle - the Higgs boson. The quark and lepton masses are generated through the gauge invariant Yukawa terms between the fermionic and scalar fields.

All in all there are 19 parameters that are not fixed by the theory, but fitted from experiments, with the final one recently determined from the discovery of the Higgs boson at the LHC [13, 14]. Yet, the predictions of the theory are consistent with all collider experimental results spanning many decades. However, the SM cannot provide an answer to some very ubiquitous observations. For example gravity has

not been worked out within the SM framework. Moreover, nothing in the theory accounts for the accelerated expansion of the universe and the seeming prevalence of non-baryonic dark matter [15]. The model is fairly symmetric between matter and anti-matter, with the exclusion of the CP-violating phase in the CKM matrix [16], so the abundance of matter is still a puzzle. It is now well established that neutrinos oscillate [17], which can only happen if they have different masses. In the SM they do not get a Yukawa coupling, so they remain massless.

In order to find explanations for these observations, we try to find deviations from the SM in particle collisions in order to gain more hints of what the true theory is. Resonances of new particles would be the most clear sign of such deviations, but discrepancy between rates of certain selection channels and their SM predicted values can also provide clues. To reach the high energy frontier of the LHC, the only viable option is to collide protons. They are composite objects, bound by the strong nuclear force. Therefore, not only is the cross section dominated by QCD background, but even on an event-by-event level QCD effects play a huge role in the final distribution of particles. For example, the reconstructed resonance of a heavy particle that decays hadronically can be washed away by the inclusion of radiation from different sources, or by a loss of energy through QCD emissions. Special techniques need to be developed to circumvent these difficulties. This thesis proposes and shows the use of such techniques for analysing various final states. In Chapter 2, the method of shower deconstruction [5,18,19] is employed to distinguish between quark- and gluon-initiated jets - a central, but still not a concluded, topic to QCD phenomenology from the conception of the theory. Two simple examples of the use of such a tagger are shown. In Chapter 3 we use very recent advances in Monte Carlo simulations of collinear electroweak boson emissions to reconstruct a hadronic W, emitted in the vicinity of a boosted quark, by several techniques and compared to the leptonic case. Finally, Chapter 4 is dedicated to identifying the notorious and background-dominated semileptonic $t\bar{t}H(b\bar{b})$ events. Moreover, limits on the measurement of the signal strength are calculated with a simple model of the systematic uncertainties.

1.1 QCD

Quarks were introduced as the constituents of the strongly interacting hadrons in order to systematise the quantum numbers of the multiple hadronic species discovered during the mid 20^{th} century. In order to predict the pattern of the light baryons and mesons, only a set of three spin-$\frac{1}{2}$ constituents with fractional electric charge seemed sufficient [20, 21]. However, it was noted that the Δ^{++} baryon, a spin-$\frac{3}{2}$ hadron, has a symmetric wavefunction in space, spin and flavour, but it is a fermion and needs to be anti-symmetric overall. To this end, the constituent quarks are required to have another type of quantum number, colour [22, 23], with three possible values, which can be made totally anti-symmetric in the Δ^{++} wavefunction [24]. Another evidence for the three colour species of quarks comes from the ratio $R = \sigma(e^+e^- \to hadrons)/\sigma(e^+e^- \to \mu^+\mu^-)$. In the quark model, away from resonances, it should be a constant and depend on the number of quark species $R \approx N_c \sum_f Q_f^2$, with Q_f the electric charge of the quark f and N_c the number of degrees of freedom per flavour. Experimental results are consistent with three colour species for each flavour [25]. Moreover, the observable particles are all singlets in this quantum number. Therefore, the quarks are confined within the hadrons and the strong interaction has a limited range.

The evidence that the quarks are real physical objects and not simply a book-keeping tool came from deep inelastic scattering experiments [26]. The (near) independence of the cross section on the virtuality of the probing photon (Bjorken scaling [27]) suggests that the hadrons contain point electric charges within them, dubbed "partons" at the time. To exhibit this scale independence, the hadrons must also be free from internal interactions at the energy of the experiment. To reverse the argument, since the independence is only approximate, there are interactions between the quarks, but they become smaller as the hadron is probed at smaller distance. This behaviour is called "asymptotic freedom" and is a crucial observation that needs to be exhibited by a theory of the strong nuclear force. Further results about the fraction of longitudinal and transverse virtual photon absorption confirm that these partons are spin-$\frac{1}{2}$ [28]. The electrically charged partons accounted for only about half the hadron momentum [29], hinting that there might be other parton

species within the hadrons.

A theory of the strong interaction between quarks needs to explain the confinement at low energy scales and the asymptotic freedom at high energies. Moreover, it also has to predict the deviation from a free theory at each photon virtuality scale.

1.1.1 QCD Lagrangian

We know that the full spectrum of quark flavours is six and, as far as the strong force is concerned, the only difference between them is their mass. The theory that fits the experimental observations listed earlier, Quantum Chromodynamics, is based on a local SU(3) symmetry, where each flavour species of quark forms a multiplet of Dirac spinors that transforms in the fundamental representation of SU(3)

$$\psi_i \to \psi_i' = \exp\left\{ i \sum_a \alpha(x)^a \, T^a \right\}_{ij} \psi_j. \tag{1.1.1}$$

Each T^a is a generator of the group corresponding to one of the independent infinitesimal transformations and $\alpha^a(x)$ are scalar functions that parametrise how much of each independent transformation is performed. The rest of the group members can be built from these generators. The T^a matrices are normalised such that $tr[T^a T^b] = T_R \delta^{ab}$ with $T_R = 1/2$. The commutation relations between the T^a define the group,

$$\left[T^a, T^b\right] = i f^{abc} T^c \,. \tag{1.1.2}$$

The numbers f^{abc} are the structure constants and they are antisymmetric under the exchange of any two of the indices. SU(3) has 8 independent generators (9 complex components, 9 fixing conditions from unitarity $UU\dagger = \mathbb{1}$, and one from $|U| = 1$). The Casimir operators in the fundamental and adjoint representations are $C_F = 4/3$ and $C_A = 3$. Just like in QED, the requirement of local invariance of the Dirac Lagrangian density necessitates the introduction of a covariant derivative through interaction terms between conserved currents of the type $j^a_\mu = \mathrm{T}^a_{ji} \bar{\psi}_j \gamma_\mu \psi_i$ and vector fields A^a_μ, called gluons,

$$\mathcal{L}_{\text{Dirac}} = \sum_f \bar{\psi}_i \left(i \partial\!\!\!/ - m_f \right) \delta_{ij} \psi_j \to \sum_f \bar{\psi}_i \left(i D\!\!\!/ - m_f \right)_{ij} \psi_j \,. \tag{1.1.3}$$

The notation $\not{v} \equiv \gamma_\mu v^\mu$ is used, where v_μ is a four-vector and γ_μ are the Dirac matrices that satisfy $\{\gamma_\mu, \gamma_\nu\} = 2g_{\mu\nu}$, with $g_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$ - the Minkowski metric. The covariant derivative is a matrix in colour space,

$$[D_\mu]_{ij} = \partial_\mu \, \delta_{ij} - igA^a_\mu T^a_{ij} \; . \tag{1.1.4}$$

The condition of gauge covariance fixes the transformation properties of the vector fields A^a_μ. In order to satisfy $D'\psi' = \exp\{i \sum_a \alpha(x)^a \, T^a\} D\psi$, the infinitesimal transformation of the vector fields is

$$A'^a_\mu(x) = A^a_\mu(x) + \frac{1}{g}\partial_\mu \alpha^a(x) + f^{abc} A^b_\mu(x)\alpha^c(x) = A^a_\mu(x) + \frac{1}{g} D^{ab}_\mu \alpha^b(x) \; , \tag{1.1.5}$$

with the covariant derivative in the adjoint representation $(T^b)_{ac} = if^{abc}$. The gluons can also form a gauge invariant term in the Lagrangian density through the field strength tensors

$$\begin{aligned} -igF^a_{\mu\nu} T^a &= [D_\mu, D_\nu] \\ F^a_{\mu\nu} &= \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu \; . \end{aligned} \tag{1.1.6}$$

This is different from the field strength of the Abelian QED, which only has the derivative terms and is gauge invariant under U(1). Therefore, in QCD contracting a field strength tensor with itself does not give a gauge invariant term to the Lagrangian density; however, the sum over all eight such terms is invariant,

$$\mathcal{L} \supset -\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu}_a \; . \tag{1.1.7}$$

The gluon kinetic term contains contributions like AAA and $AAAA$, which correspond to interactions among the gluon fields themselves. This is not surprising given the fact that they transform into one another under gauge transformations, but it is in contrast to the Abelian QED theory, where the photon does not interact with itself. The gluon self interaction is responsible for the confinement of colour-charged particles to small distances. In contrast to QED, where two oppositely charged particles can move apart and reduce the field flux density between them, corresponding to smaller force, the gluon self interaction means that the field lines between a quark and an anti-quark form a dense tube keeping the force constant. Therefore, very

quickly the gluon field between the quarks will have enough energy to create another quark anti-quark pair and form two colour-neutral systems. There has not been an analytic proof of confinement from field theory, but lattice QCD numerical calculations have shown it to be true [30–32].

Combining the gluon kinetic term and the Dirac term, the classical Lagrangian density for a local SU(3) symmetry between the quarks is

$$\mathcal{L}_{\text{classical}} = -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a + \sum_f \bar{\psi}_i \left(i\not{D} - m_f\right)_{ij} \psi_j. \tag{1.1.8}$$

1.1.2 Perturbative QCD

Such a Lagrangian is only useful if it can lead to calculations of measurable physical observables. For a field theory the transition from one field configuration $\phi_{\text{I}}(0, \boldsymbol{x})$ to another, $\phi_{\text{F}}(t, \boldsymbol{x})$ after time t is [33]

$$\langle \phi_{\text{F}}(t, \boldsymbol{x})| \exp\{-iHt\} |\phi_{\text{I}}(0, \boldsymbol{x})\rangle = \int \mathcal{D}\phi \exp\left\{ i \int_0^t d^4x \mathcal{L} \right\}, \tag{1.1.9}$$

where H is the Hamiltonian, $\mathcal{L}$ is the Lagrangian density and $\mathcal{D}\phi$ is the infinitesimal difference between field configurations according to the path integral formalism. Here $\phi(x)$ is a schematic way of incorporating the state of all the different QCD fields - the six flavours and three colours of quarks and anti-quarks (f and i), the eight colours of gluons (a), as well as their spinor and vector indices (β and μ). Therefore, the functional integral measure is

$$\mathcal{D}\phi \equiv \left(\prod_{f,i,\beta} \mathcal{D}\bar{\psi}^\beta_{f,i} \, \mathcal{D}\psi^\beta_{f,i} \right) \left(\prod_{a,\mu} \mathcal{D}A^a_\mu \right). \tag{1.1.10}$$

The QCD Lagrangian density consists of the classical part from Eq. (1.1.8) as well as two other terms that are a consequence of gauge selection,

$$\mathcal{L} = \mathcal{L}_{\text{classical}} + \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{ghost}}\,. \tag{1.1.11}$$

In a free theory the Lagrangian density is quadratic and, therefore, the path integral is exactly calculable. In a gauge theory like QCD, there are higher order terms, which spoil the form of the exponential. The only way to reach an analytical result[1] is by

[1] Lattice QCD is an alternative way to make predictions from the QCD Lagrangian using numerical methods, but its applications are limited to low energy scale.

expanding the exponential around the free Lagrangian and only evaluate terms up to a fixed order of the coupling constant. The perturbative approximation is good when this constant is small. In this picture, the particles are created and annihilated in vertices that are defined by the interaction Lagrangian, but propagate as if they are free between these vertices. The amplitude can be read off from Feynamn diagrams, which keep track of the possible contributions to a fixed order, following a set of rules derived from the Lagrangian density. There are certain transformations that need to be implemented in the Lagrangian in order to make such objects computable. They account for the two additional Lagrangian terms.

In order to define the propagator of the gauge bosons, one has to find the inverse of the quadratic term. When the boson is massless, the inversion is not possible unless the gauge degree of freedom is removed. A popular choice of gauge-fixing term comes from the Lorentz condition, which leads with the Faddeev-Popov method [34] to $\mathcal{L} \supset \mathcal{L}_{\text{gauge}} = -\frac{1}{2\xi}(\partial^\mu A^a_\mu)^2$. In addition, there is another term from the procedure, which can be inserted into the Lagrangian density as a kinetic and interaction term of massless anti-commuting scalar fields in the adjoint representation, which are called "ghosts", $\mathcal{L} \supset \mathcal{L}_{\text{ghost}} = \bar{c}^a \partial^\mu D^{ab}_\mu c^b$. These particles are not physical because they disobey the spin-statistics relation and have a negative norm. But it is no coincidence that they show up in the gauge boson propagation definition. A physical massless vector boson has only two polarisations, but when the propagator is defined in a Lorentz invariant way, there will be four components that are propagated. The addition of diagrams, containing ghosts, removes the spurious polarisations.

There is another type of gauge fixing, called axial gauge [24], which breaks Lorentz invariance by introducing a fixed direction $\mathcal{L}_{\text{gauge}} = -\frac{1}{2\xi}(n^\mu A^a_\mu)^2$. Such a gauge removes the need for ghost particles. In general however, such a term introduces a term in the propagator proportional to $(n{\cdot}p)^{-1}$, which diverges when the momentum is proportional to n_μ. It is useful when dealing with collinear singularities because the possible momentum vectors are constrained, so n_μ can always be chosen to point away from p_μ.

When perturbation theory is used to calculate transition amplitudes, ultraviolet divergences appear from the integration of loop momenta. For a renormalisable

theory, such as QCD, it is possible to systematically remove these divergences by re-defining the parameters of the theory and in the process introducing counter terms to the interaction Lagrangian. There is a freedom to choose from what scale these counter terms subtract the divergences from the bare parameters of the theory, which will affect the Feynman rules and consequently the results for a calculation at a fixed order in perturbation theory. The parameters of the theory become dependent on the renormalisation scale μ^2. In particular, the dependence of the renormalised coupling $\alpha_s = \frac{g^2}{4\pi}$ satisfies the differential equation

$$\mu^2 \frac{\partial \alpha_s}{\partial \mu^2} = \beta(\alpha_s)\,, \tag{1.1.12}$$

where the function $\beta(\alpha_s)$ can be expanded in orders of α_s, $\beta(\alpha_s) = -b_0\alpha_s^2 + \mathcal{O}(\alpha_s^3)$. The constant $b_0 = (11C_A - 2N_f)/12\pi$ can be calculated by applying the Callan-Symanzik equations for a set of Green's functions [35, 36]. Assuming that the coupling constant is small at two scales μ^2 and Q^2, its magnitudes at those points are related by

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + b_0\alpha_s(\mu^2)\log\frac{Q^2}{\mu^2}}\,. \tag{1.1.13}$$

If b_0 is positive, which is the case for QCD with six flavours of quarks, the strength of the coupling is reduced as the scale of the interaction increases. This confirms the property of asymptotic freedom and justifies the use of perturbation theory at high scale.

1.2 Experimental Setup

The focus of the thesis is exclusively on the LHC experiment. As already pointed out, this is a proton-proton collider. The two general-purpose detectors, CMS [37] and ATLAS [38], naturally have different designs, but the general layout is similar, see Fig. 1.1.

Closest to the beam axis is a tracker system, which can trace the path of charged particles. Among other applications, this allows the reconstruction of vertices, which helps in separating the radiation from collisions of different pairs of protons within the same bunch. This type of contamination is called Pile Up (PU). Moreover, the

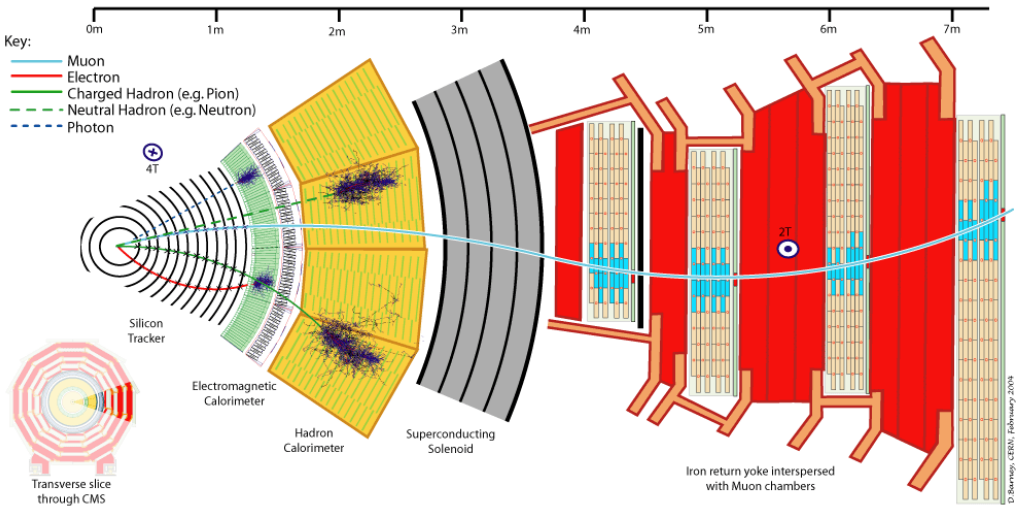


Figure 1.1: Transverse slice of the central region in the CMS detector [4]

B-mesons lifetimes are long enough that a displaced vertex can be traced, which allows for a very accurate b-tagging.

The next layer is an electromagnetic calorimeter. As the name suggests, it is able to contain the energy from electrons, positrons and photons, thereby allowing the identification of such particles. It is not straightforward to identify EM objects from the hard matrix element since many hadrons may decay to leptons or photons and produce the same signature. Therefore, it is crucial to require an isolation criterion for the hadronic radiation around and above the supercluster in the ECAL that contains the photon or electron candidate. The experiments use a Boosted Decision Tree (BDT) classifier that incorporates many observables including the isolation criterion [39] when an electromagnetic particle is identified. The difference between photons and electrons comes from the presence or lack of a charged track that leads to the ECAL supercluster. The electromagnetic calorimeter is encompassed by a hadronic calorimeter, whose granularity is significantly worse. The hadronic objects that are usually employed in an analysis are clusters of HCAL towers, called jets. They will be defined later in this chapter.

Finally, the last layer is the muon detector, which is similar in effect to the tracker, because it traces the path of the muons. The layers are subjected to magnetic fields that curve the path of the charged particles.

Apart from the PU radiation, which is the consequence of multiple proton colli-

sions, there are other contributions to the pollution of the final state. Even though the main collision is between one parton from each proton, the rest of the partons within the two colliding protons also interact and can produce detectable radiation, which is called Underlying Event (UE). This is different from PU because radiation from the UE will be traced to the same vertex as the hard interaction. Moreover, even the two partons that produce the hard collision may undergo radiation of a particle at a resolvable scale before entering into the hard vertex.

Not knowing the exact momentum fraction of the partons means that the lab frame is not the centre of mass frame of the hard collision. Nevertheless, we know it is boosted along the beam axis. Therefore, it is helpful to use quantities that are Lorentz invariant under such a boost. It is customary to assign the coordinate system such that the z-axis follows the beam, the x-axis points towards the centre of the ring and the y-axis point upwards. Then the component of the momentum in the $x-y$ plane, the transverse momentum p_T, is invariant. Also, as the system begins with no transverse momentum, the sum of transverse components in the final state should cancel. An event with a large overall p_T is indicative of an invisible to the detectors particle, such as a neutrino or a BSM particle. Another invariant quantity is the azimuthal angle ϕ in the plane transverse to the beam. The polar angle θ is affected by the boost and the relative polar angles between objects in the lab frame is different from what would be measured in the c.o.m frame. More useful variables are the rapidity y and pseudorapidity η

$$y = \frac{1}{2}\log\frac{E+p_z}{E-p_z}\ , \quad \eta = -\log\theta/2\ . \tag{1.2.14}$$

The rapidity depends on the frame of reference, but the difference between the rapidity of two objects is constant under a boost in the z direction. For massless particles $y = \eta$. So for the final state particles that reach the detectors, the pseudorapidity is a good approximation. Its benefit is that it is purely geometrical. Therefore, an angular distance between objects can be defined as $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$.

1.3 From perturbative Matrix Element to observable Cross Section

In collision experiments one can observe the rate at which a final state of interest occurs given an experimentally controlled initial state. Even though the number of transitions is related to the transition amplitude between the "in" and "out" states, there are experiment-dependent contributions as well. For example, if two beams are fired at each other, the number of collisions will depend on the number of all participants. It will, therefore, be proportional to the number densities in both beams, the velocity of the beams, the time of beam overlap, and the area of overlap. All of these quantities enter in the experiment-dependent integrated luminosity $\int Ldt$. For a description of the integrated luminosity at the LHC check [40]. The number of events is proportional to this quantity, $N = \sigma \int Ldt$. The constant of proportionality σ, the cross section, is independent of the experimental set-up and determined by the underlying physics; therefore, it can be compared to theory.

The transition amplitude for $2 \to n$ collision is ${}_{\text{out}}\langle p_1, p_2, ..., p_n | k_1, k_2 \rangle_{\text{in}} = \langle p_1, ..., p_n | S | k_1, k_2 \rangle$. The variables p_i, for $i \in [1, n]$, are the four-momenta of the outgoing particles and, analogously, k_1 and k_2 are those of the incoming particles. Since we are not interested in the transitions that do not change the initial state, we remove the unit part of the S matrix $S = \mathbb{1} + iT$. Then the invariant matrix element is defined by [33]

$$\langle p_1, ..., p_n | iT | k_1, k_2 \rangle = (2\pi)^2 \delta^{(4)}(k_1 + k_2 - p_1 - ... - p_n) i\mathcal{M}(k_1, k_2 \to p_1, ..., p_n) \, . \tag{1.3.15}$$

An overall momentum-conserving delta function is factored out of $\mathcal{M}$. The matrix element can be calculated to a fixed order in α_s by summing the contributions of all connected and amputated Feynman diagrams that match the external particles, using the rules and parameters of the renormalised theory [33, 41]. The cross section

is then a combination of the dynamics and the kinematics of the process

$$\sigma_{2\to n} = \frac{1}{2k_1^0\ 2k_2^0\ |v_1 - v_2|} \int \left(\prod_{i=1}^{n} \frac{d^4p_i}{(2\pi)^4} \delta(p_i^2 - m_i^2) \right) \times |\mathcal{M}(k_1, k_2, p_1, ..., p_n)|^2\ (2\pi)^4 \delta^{(4)}(k_1 + k_2 - p_1 - ... - p_n)\ . \tag{1.3.16}$$

1.3.1 Infrared and collinear divergence

The Infrared problem

In renormalised perturbation theory the UV divergences are absorbed in the parameters of the Lagrangian and the cross section depends on the renormalisation scale through the running of the coupling and the masses of the matter fields. But there are other divergences in the matrix elements calculated at fixed order, which occur at the other end of the energy spectrum. These divergences are most easily illustrated by the cross section for the emission of a soft gluon from a quark anti-quark pair in the final state. The matrix element is the sum of the two Feynman diagrams (Fig. 1.2): when the gluon with momentum k is emitted from the quark (momentum p) and when the gluon is emitted from the anti-quark (momentum p') [33]

$$\begin{aligned} [\mathcal{M}_{q\bar{q}g}]^a_{ij} &= \bar{u}(p)\ igT^a_{ij} \not{\epsilon}(k) \frac{i(\not{p} + \not{k} + m)}{(p+k)^2 - m^2}\ \mathrm{M}v(p') \\ &+ \bar{u}(p)\mathrm{M}\ igT^a_{ij} \frac{-i(\not{p}' + \not{k} - m)}{(p'+k)^2 - m^2} \not{\epsilon}(k)\ v(p') \\ &\approx [\bar{u}(p)\mathrm{M}v(p')]\ \ gT^a_{ij} \left(\frac{p \cdot \epsilon(k)}{p \cdot k} - \frac{p' \cdot \epsilon(k)}{p' \cdot k} \right)\ . \end{aligned} \tag{1.3.17}$$

Here M is the rest of the diagram that couples to the quark anti-quark pair and the gluon through two Dirac spinor indices. To reach the last line, we ignore the $\not{k}$ term in the numerator of the propagator and we apply Dirac's equation for the outgoing fermions. In order to find the cross section, we first need the squared matrix element, summed and averaged over the possible colours and polarisations of the final and initial particles

$$|\mathcal{M}_{q\bar{q}g}|^2 = |\mathcal{M}_{q\bar{q}}|^2\ C_F 4\pi\alpha_s \frac{2p \cdot p'}{p \cdot k\, p' \cdot k}\ . \tag{1.3.18}$$

Then we can split the phase space element for the $q\bar{q}$ pair from the one-particle phase space and completely factorise the $q\bar{q}$ cross section from the splitting of the

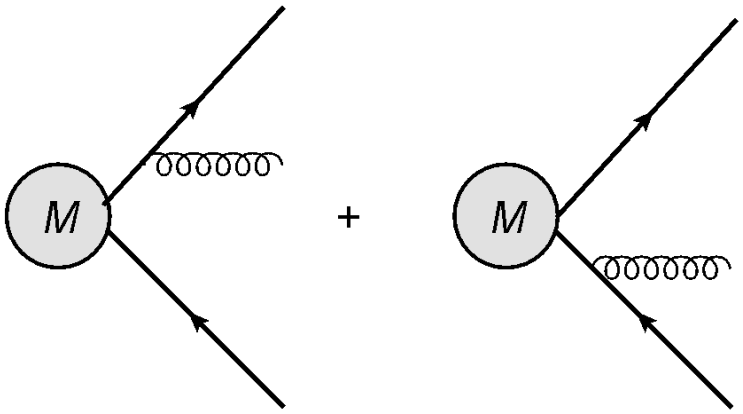


Figure 1.2: The two Feynman diagrams, contributing to an emission of a gluon from a quark anti-quark pair

soft gluon

$$\sigma_{q\bar{q}g} = F \cdot d\Phi_{q\bar{q}g} \cdot |\mathcal{M}_{q\bar{q}g}|^2 = F \cdot d\Phi_{q\bar{q}} \cdot |\mathcal{M}_{q\bar{q}}|^2 \times dE_k \, d\cos\theta \, d\phi \, \frac{\alpha_s C_F}{\pi^2} \frac{1}{E_k^2(1-\cos^2\theta)} \, . \tag{1.3.19}$$

The emission of the gluon factorises from the rest of the cross section and there is a classical probability associated with it. Moreover, this probability explodes when the gluon energy approaches zero, $E_k \to 0$, or when its momentum becomes collinear with one of the quarks, $\theta \to 0, \pi$. This is not a physical result, but also no physical measurement can look at the energy and angular distribution of the gluon with infinite accuracy. Therefore, to calculate the expected cross section from an experiment, the virtual contributions from the interference of the Born matrix element and the loop correction need to be added for a consistent perturbative calculation because they are at the same order in α_s. According to the Block-Nordsieck [42] and Kinoshita-Lee-Nauenberg theorems [43,44] these divergences cancel to all orders for inclusive cross sections, thus preserving unitarity.

In order for such cancellations to occur for pQCD predictions of other observables than inclusive cross sections, these variables need to have the property of infrared and collinear safety. This property is characterised by the following condition for observable $O(\{p\})$

$$O(p_1, ..., p_i, p_j, ..., p_{n+1}) \to O(p_1, ..., p_i + p_j, ..., p_{n+1}) \, , \tag{1.3.20}$$

if p_i and p_j become collinear to one another or either of them becomes soft. When this is true, the observable is the same for the virtual and real corrections in the enhanced region and can be taken out of the integrals over the phase space of the collinear and soft particles, allowing for the infinities in the cross section contributions to cancel exactly.

Going back to measuring $\sigma_{q\bar{q}g}$, as it stands it is not a well defined observable. Moreover, the detectors of collider experiments do not observe partons in the first place, but cascades of hadrons. Therefore, in order to compare theoretical calculations to experimental observation, we need to cluster the final state particles (be it at parton level or hadron level) into IRC safe jets. Then the cross section for N jets will depend on the jet definition, but it will be well defined as long as the jet definition is.

Sequential clustering algorithms

Such jet definitions are the family of sequential clustering algorithms. A sequential clustering algorithm combines final state objects, which for the purposes of jet clustering we refer to as seeds, a pair at a time. At each stage of the clustering, there is a distance measure d_{ij} between each pair of seeds and in the hadron-hadron collision case another distance measure d_i, defined for each seed individually. If the smallest measure is d_{ij} then seeds i and j are merged into a single seed. All distances involving the old two seeds are dropped and new ones are calculated for the combined seed. If a d_i distance is the smallest then this seed is assigned the status of a jet and is removed from further clustering. The steps repeat until all the seeds are merged and assigned to jets. The distance measures are

$$\begin{aligned} d_{ij} &= \min(p_{Ti}^{2a}, p_{Tj}^{2a}) \frac{\Delta R_{ij}^2}{R_{\text{jet}}^2} \ ; \\ d_i &= p_{Ti}^{2a} \ . \end{aligned} \tag{1.3.21}$$

The geometric distance is $\Delta R_{ij}^2 = (\phi_i - \phi_j)^2 + (\eta_i - \eta_j)^2$, where ϕ_i is the azimuthal angle around the axis of the beam and η_i is the pseudorapidity of seed i. This quantity is invariant for massless particles under boosts along the z-axis and is therefore a suitable measure for proton-proton collisions, where the centre of mass

of the hard interaction can be boosted along the beam direction in the lab frame. The other geometric parameter R_{jet}, an arbitrary choice for each analysis, controls the angular size of the jets, but is usually not exactly a jet radius in the mathematical sense. The last parameter a controls what type of seed pairs are to be clustered first. Another cut is applied on the minimum transverse momentum that a jet must have, which limits the final number of jets considerably. There are two reasons for a minimum p_T requirement. The theoretical reason is that there must be such a cut-off in order to make the algorithm IR-safe; moreover, this value should be chosen such that α_s is small at that scale. The practical reason, and the leading one, has experimental considerations. As a proton-proton interaction involves multiple parton interaction at the same time, there are multiple jets with low p_T that are not part of the hard process. A sufficiently large cut will ameliorate this source of systematic effects.

When $a = 1$, priority is given to pairs with seeds that are close geometrically and at least one of them is soft. This algorithm is called k_T-algorithm [45]. The intention is to mimic the QCD infrared and collinear singularities. A practical problem with this definition is the amorphous area that the jets take. Even though the jet definition involves a R_{jet} parameter, it determines how far away a seed needs to be from all others in order to be assigned as a jet and is not related to the area of the jet directly. This leads to difficulties in calibrating the properties of the jets according to detector effects.

A solution to this problem is to change the distance definition with $a = -1$. This algorithm is called anti-k_T [46]. It gives priority to nearby pairs with at least one seed with large p_T. Therefore a hard seed will accrete the softer radiation around it until all seeds within a radius R_{jet} are merged into it. This will form a geometrically well defined circular jet, where R_{jet} is indeed indicative of the jet size. This situation is only true for the hardest seed in the vicinity. If two hard seeds are separated by more than R_{jet} but less than $2R_{\text{jet}}$, the harder seed will form a circular jet around it, but this will come at the expense of the softer seed.

A third popular choice is $a = 0$ named the Cambridge/Aachen algorithm [47]. The distance is purely geometrical in this case. The effect of that is to soften

the "winner-takes-all" scenario with overlapping anti-k_T jets by assigning seeds to whichever jet centre is closest, disregarding the relative hardness of the competing jets.

Final state collinear factorisation

The equation in the soft limit Eq. (1.3.19) shows the factorisation of the matrix element that involves a soft or soft and collinear final state gluon into a hard matrix element and a gluon emission probability from a colour dipole. There is also similar factorisation when a nearly on-shell parton splits in two collinear partons with more symmetric distribution of the parent's energy. If the particles are massless compared to the momentum involved, then this configuration is strongly enhanced by the denominator of the propagator. When particle a splits into $a \to b + c$, the matrix element splits into [24, 33]

$$\mathcal{M}_{n+1} \to M^s \frac{\sum_\lambda (\lambda^s(a)\lambda^{s\prime}(a))}{(p_b + p_c)^2}\, Tg\, V^{s\prime}(\lambda(b), \lambda(c), p_b, p_c). \tag{1.3.22}$$

Where M^s is the hard part of the matrix element that is attached to an on-shell propagator and has a polarisation index s (either spinor or four-vector index depending on particle a). The sum is over all polarisations that are propagated through a (in the physical gauge these are only the physical polarisations). The denominator $(p_b + p_c)^2 \approx z(1-z)E_a^2\theta^2$ becomes small when the emission is collinear. Here $z = \frac{E_b}{E_a}$ and $1 - z = \frac{E_c}{E_a}$. Finally the last term $V^{s\prime}$ is the splitting vertex Feynman rule coupled with the polarisations of c and b with the colour factor and coupling constant extracted out. It also has a polarisation index to link with the propagator of a. The QCD rules that form $V^{s\prime}(\lambda(b), \lambda(c), p_b, p_c)$ are such that if the spin/polarisation of particles b and c are fixed, so is for particle a. Therefore, only one term in the sum over the propagator polarisations contributes to a non-zero result. Therefore the matrix element squared is factorisable

$$|\mathcal{M}_{n+1}|^2 \propto |\mathcal{M}_n|^2 \frac{\alpha_s}{p_b \cdot p_c} P(z, \lambda_b, \lambda_c). \tag{1.3.23}$$

If the ϕ dependence is integrated out and the contribution from all b and c polarisations is summed, then the spin averaged splitting functions are obtained [24, 33, 48]

$$\begin{aligned} P(z)_{q\leftarrow q} &= C_F \left(\frac{1+z^2}{1-z} \right) \\ P(z)_{g\leftarrow q} &= C_F \left(\frac{1+(1-z)^2}{z} \right) \\ P(z)_{q\leftarrow g} &= T_R \left(z^2 + (1-z)^2 \right) \\ P(z)_{g\leftarrow g} &= C_A \left(\frac{1-z}{z} + \frac{z}{1-z} + z(1-z) \right). \end{aligned} \tag{1.3.24}$$

The arrow notation in the subscript, $P(z)_{j\leftarrow i}$, indicates that a particle i splits and one of the daughters, a particle j, takes a fraction z of its energy. That is why $P(z)_{q\leftarrow q} = P(1-z)_{g\leftarrow q}$, they come from the same vertex $q \to qg$ but refer to the quark and gluon daughters respectively. The soft singularity is always associated with a gluon in the final state. For example $P(z)_{q\leftarrow g}$, which has only quarks, does not diverge when z or $1-z$ tends to zero.

Hadronic cross section and initial state radiation

In the context of the parton model, a hadron is treated as a collection of non-interacting partons [27,49]. Each kind of parton is associated with a number density $q(x)$, called parton distribution function (pdf), where x is the longitudinal momentum fraction of the parton. These pdf's are specific to each hadron. Then the interaction between two hadrons is the incoherent sum of the interaction of the component partons, weighted by their number density

$$\sigma_{pp\to X} = \int_0^1 dx_1 q_{a/p}(x_1) \int_0^1 dx_2 q_{b/p}(x_2) \hat{\sigma}_{ab\to X}(x_1 p_1, x_2 p_2). \tag{1.3.25}$$

Here $\hat{\sigma}$ designates the partonic cross section from the hard interaction, which can be evaluated from Eq. (1.3.16) using the Feynman rules of the theory. The pdf's are not calculable from perturbation theory because they contain the effects of QCD in the large α_s regime. They can be extracted experimentally though. Deep inelastic scattering experiments, where a lepton is used to probe the structure of a hadron, have been particularly useful in this regard [26,50]. Interestingly, the quark distributions can only account for 50% of the total hadronic momentum, therefore gluons play

an important part in hadron collisions. However, the presence of gluons requires the use of QCD to higher orders in α_s. The NLO contribution to the process is an emission of a gluon from the initial state quark line and also the 1-loop correction to the quark line. The sum of the real and virtual first order corrections is

$$\hat{\sigma}_R^{(1)} + \hat{\sigma}_V^{(1)} \propto \int \frac{d\mu^2}{\mu^2} \int dz P_{q\leftarrow q}(z)(\hat{\sigma}^{(0)}(zp) - \hat{\sigma}^{(0)}(p)). \tag{1.3.26}$$

After a real gluon emission the momentum that enters the hard scattering is zp, while a virtual gluon will keep the input momentum as the original. When the gluon is soft and $z \to 1$, the integrand approaches zero. Therefore, the soft divergences are cancelled between the real and virtual contributions as expected. However, when $z < 1$, the integral w.r.t. z is finite. But the integral over μ^2 diverges, so the partonic cross section is not an IRC-safe observable. The divergence signals the breaking of pQCD when the scale of the gluon splitting is under the perturbative regime. This divergence is ameliorated by the introduction of the parton distribution functions, which have been measured at a fixed perturbative scale and absorb all the contributions from collinear splittings. Thus, the hadronic cross section $\sigma = \sigma^{(0)} + \sigma^{(1)} = q_1(\mu_F^2) \otimes q_2(\mu_F^2) \otimes (\hat{\sigma}^{(0)} + \hat{\sigma}^{(1)})$ is IRC-safe. However, the μ^2 integral in $\sigma^{(1)}$ is still large when the scale at which the pdf's are evaluated is much lower than the scale of the partonic interaction. Therefore, $\sigma^{(1)} \propto \alpha_s(Q^2) \log \frac{Q^2}{\mu_F^2}$, explicitly shows the logarithmic dependence of the hadronic cross section on the hard scale Q^2 and confirms the observed breaking of Bjorken scaling. Moreover, the correction to the zeroth-order cross section may no longer be small if μ_F^2 is small.

These large logarithms can be absorbed into the pdf's if one could evaluate them at the scale of the hard interaction. Even though the pdf's themselves are not calculable from first principle, their evolution can be determined from perturbation theory as long as the end points of the evolution are at sufficiently large scales. The probability that a parton is going to split at a scale between μ^2 and $\mu^2 + \Delta\mu^2$ is

$$d\mathcal{P}(\mu^2) = \frac{\alpha_s}{2\pi} \frac{d\mu^2}{\mu^2} \int dz P(z). \tag{1.3.27}$$

The change in the distribution $q(x, \mu^2)$ from scale μ^2 to $\mu^2 + \Delta\mu^2$ is the difference between all the possible splittings at scale μ^2 that can lead from $x' > x$ to x and all

possible splittings at scale μ^2 that remove lower x [24]

$$\begin{aligned} q(x,\mu^2+\Delta\mu^2)-q(x,\mu^2) &= \frac{\Delta\mu^2}{\mu^2}\frac{\alpha_s}{2\pi}\int_x^1 dx' P(x')q(x',\mu^2) - \frac{\Delta\mu^2}{\mu^2}\frac{\alpha_s}{2\pi}\int_0^1 dz P(z)q(x,\mu^2) \\ &= \frac{\Delta\mu^2}{\mu^2}\frac{\alpha_s}{2\pi}\int_0^1 dz P(z)\left(\frac{q(x/z,\mu^2)}{z}-q(x,\mu^2)\right) \\ &\equiv \frac{\Delta\mu^2}{\mu^2}\frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z}P(z)_+ q(x/z,\mu^2). \end{aligned} \tag{1.3.28}$$

Where the regularised splitting functions $P(z)_+$ are defined such that $\int_0^1 dx f(x)_+ g(x) \equiv \int_0^1 dx f(x)(g(x)-g(1))$. Accounting for the different parton species one arrives at the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) evolution equations [48, 51, 52]

$$\mu^2\frac{\partial q_i(x,\mu^2)}{\partial\mu^2} = \sum_j \frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z}P_{i\leftarrow j}(z)_+ q_j(x/z,\mu^2). \tag{1.3.29}$$

Parton Shower

The Sudakov form factor $\Delta(\mu_2^2,\mu_1^2)$ is an exponential factor that dampens the cross section for observables, when they are evaluated for configurations that produce large logarithmic corrections, such as fixed number of jets cross section with a definition that is able to distinguish soft and collinear emissions. The factor is the result of summing the virtual contributions with the real emissions under the resolution scale to all orders in α_s. Assuming that unitarity is preserved, one can express the probability that an emission will not occur at a scale μ^2 to $\mu^2+\delta\mu^2$ as the negative of the probability for the emission to happen in that element (Eq. 1.3.27) and the probability that it has not happened before [24]

$$\begin{aligned} d\Delta(\mu^2,\mu_0^2) &= -d\mathcal{P}(\mu^2)\cdot\Delta(\mu^2,\mu_0^2) = -\frac{\alpha_s}{2\pi}\frac{d\mu^2}{\mu^2}\int dz P(z)\cdot\Delta(\mu^2,\mu_0^2)\ , \\ \Delta(\mu_2^2,\mu_1^2) &= \exp\left\{-\int_{\mu_1^2}^{\mu_2^2}\frac{d\mu^2}{\mu^2}\frac{\alpha_s}{2\pi}\int dz P(z)\right\}. \end{aligned} \tag{1.3.30}$$

The Sudakov factor $\Delta(\mu_2^2, \mu_1^2)$ is the probability that no resolvable emission will happen between the scales μ_2^2 and μ_1^2. The lower scale cannot drop further than a limit where pQCD is no longer valid.

Then it is possible to generate soft and collinear emissions according to these distributions from the hard interaction scale down to the cut-off scale using Monte Carlo methods [24]. In order to generate an emission, three numbers need to be generated - the scale at which it happens, the momentum fraction of the emitted particle and its azimuthal angle. If we start at a large scale μ_2^2, then we select the scale of the emission μ_1^2 according to $\Delta(\mu_2^2, \mu_1^2) = R_1$, where R_1 is a random number from the uniform distribution $[0, 1]$. It is possible to generate a low R_1, such that μ_1^2 is less than the cut-off. In that case nothing is emitted from the current branch below μ_2^2. Alternatively, if the new scale is legitimate, then another random number is selected to solve the following for z

$$\int_{z_{\min}}^{z} dz' \frac{\alpha_s}{2\pi} P(z') = R_2 \int_{z_{\min}}^{1-z_{\min}} dz' \frac{\alpha_s}{2\pi} P(z'). \tag{1.3.31}$$

The random number distribution is again uniform between $[0, 1]$. The limit on the z integration is determined from the limit imposed by the cut-off scale. Now that the scale of the splitting and the momentum fraction are known, the only thing left to generate is the azimuthal direction of the decay plane by choosing another random number from a uniform distribution $[0, 2\pi]$. This procedure is repeated to the two new branches with the only difference of replacing $\mu_2^2 \to \mu_1^2$.

Such a probabilistic evolution is used by event generators, such as Pythia [53], Herwig++ [54], Sherpa [55], to bring calculations at fixed order matrix elements and hard scale down to the limits of perturbation theory. In order to fully simulate the final state of hadronic collisions, they also implement data-driven models of the hadronisation of the partons from the shower. Moreover, they implement underlying event radiation models. This makes Monte Carlo event generators an indispensable (and only) tool for bridging theoretical calculations of high energy particles with collision experiments final states.

Chapter 2

Shower deconstruction for quark-gluon tagging

Ever since the first collider experiments, the hadronic radiation has been associated, through jet definitions, with gluons or quarks, produced in the hard interaction. Even though there has been a great deal of understanding from the advent of QCD of the difference between quark and gluon originating jets, such as the average multiplicity [24, 56] or broadening of jets [57, 58], as a result of the colour charges, we still treat a jet indiscriminately as either quark or gluon. We are unable to separate events based on quark and gluon jets, so we are forced to treat all the same, increasing the background that experimentalists need to control. Being able to tag quarks and gluons on a jet-by-jet basis will go a long way in reducing QCD background. To date it is only viable to tag jets originating from b quarks thanks to the displaced vertex of the B-meson decay, despite efforts to find variables that are useful in separating light quarks from gluons [58–64], which have been studied by ATLAS [65] and CMS [66].

Two examples of specific searches, which can benefit from quark and gluon tagging, and which will be used as a showcase at the end of the chapter, are the recoil of a mediator that decays invisibly to dark matter particles from a single jet [67] and the weak boson fusion production of the Higgs boson [68–71]. In the first case, given a scalar Higgs-like mediator that couples to the top quark, the recoiling jet will originate with comparable rate from a gluon or a quark. The mono-jet background from

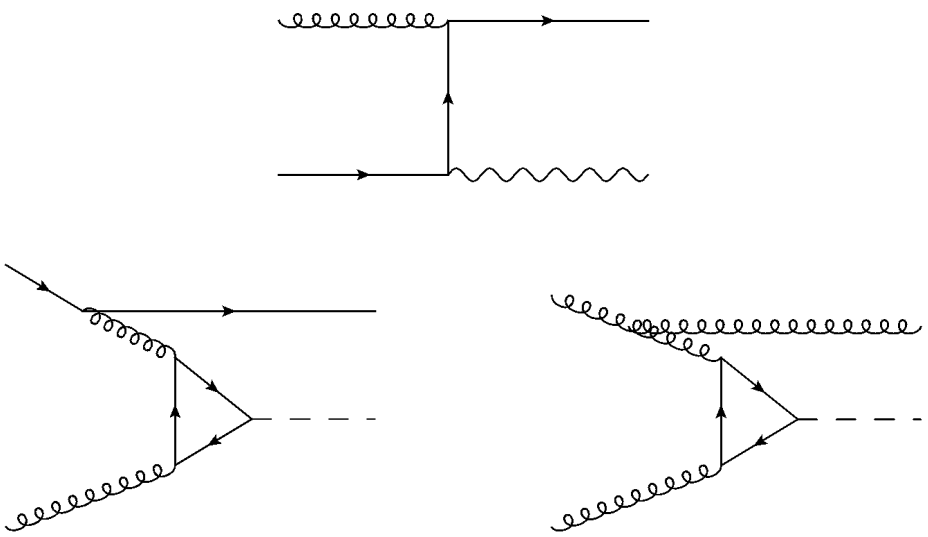


Figure 2.1: Top: the main background to dark matter mono-jet search from $qg \to qZ(\nu\bar{\nu})$. Bottom: production of a scalar dark matter mediator in association with a quark (left) and a gluon (right).

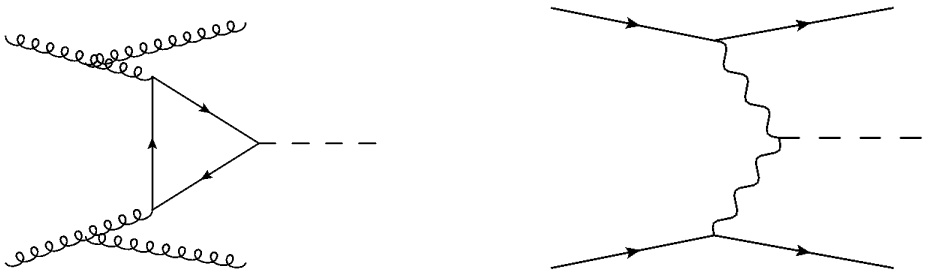


Figure 2.2: Left: Higgs boson in association with two jets production through gluon fusion. Right: Higgs boson production through weak boson fusion.

the SM is an invisible $Z(\nu\bar{\nu})$ boson with predominantly a quark recoil (see Fig. 2.1). In the second example, the signal event is expected to contain two quark-initiated jets, but the dominant Higgs production mode is through gluon fusion, which can mimic the topology of WBF events with gluon-initiated jets [72] (see Fig. 2.2).

In this chapter we propose an additional variable based on Shower Deconstruction that provides a competitive quark and gluon purification and is IRC-safe as well as robust to experimental systematic effects.

2.1 Shower deconstruction

2.1.1 The most sensitive tagger

Any tagging procedure or analysis in high energy particle physics is an attempt to take the space of the energies and directions of all final state particles (or even detector read-outs) and somehow map the distributions from interesting and uninteresting events into a single test statistic in a way that preserves enough of the discrepancy between the different types of events. An event shape for example uses the momenta of all reconstructed particles/topoclusters and compresses all that information into a single number. Usually the probability distribution over the original phase space is transformed through jet algorithms or grooming techniques to reduce the dimensionality or shift signal-rich bins away from background-rich bins. During the compression process, inevitably a lot of information is lost and some of that information is important for separation of the hypotheses. Therefore, a bad choice of mapping can compromise the sensitivity. However, the reverse is not true; a good mapping cannot improve the sensitivity to 100% efficiency as long as there is an overlap between the distributions in the original space. It is unreasonable to expect that, if two hypotheses produce similar probability distributions under the most fine-grained measurement, we could define a function over these variables that will enlarge the discrepancy between the hypotheses as the probability is reassigned to the new variable. There is therefore a best test statistic. By the Neyman-Pearson lemma [73] the test statistic that rejects the most amount of background by keeping a fixed efficiency is the likelihood ratio. A region C_{LR} in the space of the independent variables x defined by a cut on the likelihood ratio c that leaves ϵ_0 signal efficiency is

$$\begin{aligned} C_{LR} &= \left\{ x : \frac{L(S|x)}{L(B|x)} \geq c \right\} \\ P(x \in C_{LR}|S) &= \int\limits_{C_{LR}} L(S|x)dx = \epsilon_0 \ , \end{aligned} \tag{2.1.1}$$

and the probability of the background events that fall in the same cut is

$$P(x \in C_{LR}|B) = \int\limits_{C_{LR}} L(B|x)dx \ .$$

In the context of particle collision events, the variables x could be anything from the components of the momenta of each final state particle to a single variable, like the leading jet mass. If we choose a generic region defined by the cut C_G in the same space of variables x such that the probability of the signal events that are kept is unchanged, $P(x \in C_G|S) = \epsilon_0$, then the probability of background events $P(x \in C_{LR}|B) \leq P(x \in C_G|B)$ according to the lemma. To see this consider the difference between the probabilities in the two regions. The two regions can be separated into three subregions of interest

$$C_{LR} \cup C_G = C_{LR} \cap C_G + C_{LR} \cap C_G^c + C_{LR}^c \cap C_G \ , \tag{2.1.2}$$

where S^c is the complement of set S. The first subregion is common to both, so the difference in background probability is not going to come from there. Therefore in order for the lemma to be true, $P(x \in C_{LR} \cap C_G^c|B) \leq P(x \in C_{LR}^c \cap C_G|B)$. The left hand side is a probability defined over the region C_{LR}, where the condition $\frac{L(S|x)}{L(B|x)} \geq c$ holds, leading to

$$\begin{aligned} P(x \in C_{LR} \cap C_G^c|B) = \int\limits_{C_{LR} \cap C_G^c} L(B|x)dx \leq \\ \frac{1}{c} \int\limits_{C_{LR} \cap C_G^c} L(S|x)dx = \frac{1}{c} P(x \in C_{LR} \cap C_G^c|S) \ . \end{aligned} \tag{2.1.3}$$

According to the condition that the signal efficiency is constant in both regions C_{LR} and C_G, the signal probability in the non-overlapping subregions must be equal,

$$P(x \in C_{LR}|S) = P(x \in C_G|S) \Rightarrow P(x \in C_{LR} \cap C_G^c|S) = P(x \in C_{LR}^c \cap C_G|S) \ ; \tag{2.1.4}$$

therefore, $P(x \in C_{LR} \cap C_G^c|S) = P(x \in C_{LR}^c \cap C_G|S)$. The region on the right hand side is in the complement set of C_{LR} and so the reverse condition holds there, i.e. $\frac{L(S|x)}{L(B|x)} \leq c$. Thus,

$$\begin{aligned} \frac{1}{c} P(x \in C_{LR} \cap C_G^c|S) = \frac{1}{c} P(x \in C_{LR}^c \cap C_G|S) \\ \leq \int\limits_{C_{LR}^c \cap C_G} L(B|x)dx = P(x \in C_{LR}^c \cap C_G|B) \ , \end{aligned}$$

which results in the inequality

$$P(x \in C_{LR}|B) \leq P(x \in C_G|B) \ . \tag{2.1.5}$$

Therefore, for a fixed signal efficiency, the likelihood ratio keeps the least amount of background of all possible test statistics. Essentially the proof shows that for any variable with contours that are not parallel to the likelihood ratio contours in the space of x, a region with high likelihood ratio is replaced by a region with a lower likelihood ratio with the same amount of signal. Therefore, since the likelihood ratio is lower in the new region, the overall background will increase. Thereby, the S/B ratio will be lower for any cut that does not follow a likelihood ratio contour.

2.1.2 Shower deconstruction framework

The paradigm behind Shower Deconstruction [5, 18] is to look for this exact test statistic defined over as many dimensions as practically possible. The same philosophy underpins the Matrix Element Method (MEM) [74,75], in which the probability of a final state configuration is estimated from the leading order matrix elements of signal and background processes and combined into a likelihood ratio. As the complexity of the cross section estimation increases with the object multiplicity there is a practical limit on how many jets can be involved in the calculation. The conception of the shower deconstruction method originates as an attempt to improve the tagging of a boosted Higgs boson ($H \to b + \bar{b}$) compared with a QCD jet, initiated by a single gluon [5]. For a boosted particle, the subsequent decays and partonic evolution will be collinear and therefore the probability of a final state evolving from different hypothetical initiating particles is calculated in the context of the collinear approximation as opposed to fixed order matrix elements.

The shower deconstruction method begins from an anti-k_T fat jet. In order to limit the dimensions of the space over which the likelihood ratios are calculated, the constituents of the jets are reclustered into smaller $R_{\text{jet}} = 0.2$ k_T-algorithm microjets. For computational purposes only the hardest N microjets are kept, where N does not exceed 10 and is often less than that. For the original implementation of the method $N \leq 7$ seemed to provide comparable discrimination with larger limits. Therefore, a microjet configuration is defined by the set of four-momenta $\{p\}_N$. In a situation where b-quarks are involved and can be tagged, which is often the case and certainly true for $H \to b\bar{b}$ tagging, an additional discrete variable is defined for

the b-tag of the microjet. Realistically, not all 7 microjets can be b-tagged, so in the case of [5] only the hardest three microjets are. With this consideration in mind, a fat jet configuration in the framework of shower deconstruction is a set $\{p,t\}_N$ of $4N$ continuous variables and up to three boolean variables $t \in \{T,F\}$ - one for each of the hardest three microjets. This is a lot less than the degrees of freedom of $\mathcal{O}(100)$ hadrons that constitute the fat jet, but is still a sizeable space. The likelihood ratio is a function over all possible fat jet configurations,

$$\chi\left(\{p,t\}_N\right) = \frac{P\left(\{p,t\}_N|S\right)}{P\left(\{p,t\}_N|B\right)} \quad . \tag{2.1.6}$$

In principle the probability distributions can be numerically estimated using Monte Carlo parton showers such as Herwig [54], Pythia [53], or Sherpa [55]. However, even if each variable is binned very coarsely, the number of bins that need to be filled by these Monte Carlo generators is impractically large. Not to mention that once the function over all bins is estimated, it would have to be stored and accessed each time a jet tagging is requested.

The approach undertaken in shower deconstruction is to build all possible evolutions from the hypothesis particle (QCD parton or Higgs boson in the case of [5]) to the final set of microjet configuration $\{p,t\}_N$. There is a probability associated with each of these histories. The total probability for each initiating particle is the sum of the probabilities of all histories derived from it. Therefore the equation for the likelihood ratio function is transformed as

$$\chi\left(\{p,t\}_N\right) = \frac{P\left(\{p,t\}_N|S\right)}{P\left(\{p,t\}_N|B\right)} = \frac{\sum\limits_{h\in S} P\left(\{p,t\}_N|h\right)}{\sum\limits_{h\in B} P\left(\{p,t\}_N|h\right)} \quad . \tag{2.1.7}$$

It is important to stress that the histories are different for the different hypotheses. After all, by definition the history begins from a different particle in the signal and background, so histories cannot repeat between the numerator and denominator. The probability for each history is calculated analytically using a diagramatic approach similar to Feynman diagrams, but in the collinear and soft limits. Moreover, the history probability is individually computable by multiplying the expressions associated with the elements because the diagram rules refer to classical probabilities and not quantum-mechanical amplitudes. An example of a history is shown

in Fig. 2.3. It consists of $1 \to 2$ particle vertices and "propagators" that take into account the non-emission between shower times, with the additional information about the colour connected partners. The object that calculates the probability for non-resolvable emission is the Sudakov factor; therefore, the branch lines in a diagram correspond to that. This evolution time of the decay of a particle is, in accordance with the parton shower described in [76–78],

$$t = log\left(\frac{|Q_0|k_J}{\mu_J^2}\right) \quad , \tag{2.1.8}$$

where k_J and μ_J are respectively the transverse momentum and virtuality of the branch and $|Q_0|$ is the scale of the hard interaction. Each of the lines that does not end in a vertex is assigned to one of the final state microjets. The vertices, except for the hard interaction, have to be acceptable SM vertices (cannot have quark splitting to two gluons for example). The diagram in Fig. 2.3 is not symmetric under the interchange of the left and right branch coming out of a vertex. The colour connections between colour-charged particles affect the radiation pattern. Therefore, assigning colour partners is important for the decay and evolution of those particles. Keeping the relative left-right position is a good way to track those colour connections to leading colour approximation. Going back to the likelihood ratio formula, the probability from each hypothesis can now be expressed as

$$\begin{aligned} P\left(\{p,t\}_N|S\right) = \sum_{h\in S}\Bigg(&\prod_{i=1}^{N} H\left(p_i^{RB}, p_i^{LB}, p_i^{LC}, p_i^{RC}, p_i^{GM}\right) \\ &\times \prod_{j=1}^{2N} \Delta\left(p_j^{J}, p_j^{LC}, p_j^{RC}, p_j^{GM}\right) \prod_{k=1}^{N} B\left(f_k, t_k\right)\Bigg) \quad . \end{aligned} \tag{2.1.9}$$

In this equation H is the probability for each of the N splittings as a function of the particles associated with vertex i: the outgoing particles (p^{RB}, p^{LB}); the colour partners (p^{RC}, p^{LC}); the grand mother particle p^{GM}. By "a function of a particle" it is meant a function of both the four-momentum and flavour. The Δ terms are the $2N$ propagator lines, which are functions of the propagated particle p^J, the mother particle p^{GM}, and the colour connected partners (p^{RC}, p^{LC}). Finally, the B functions are the probabilities that the final state branch k will be tagged as the flavour f_k, which is assigned from the history evolution, given the b-tag t_k of the microjet it represents.

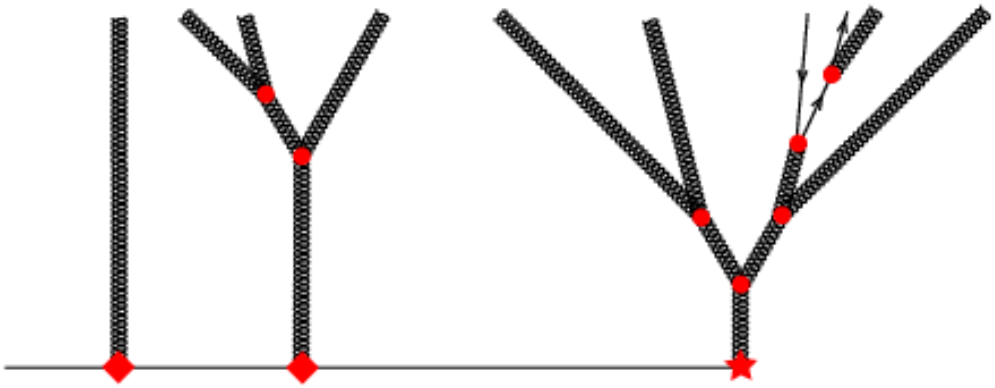


Figure 2.3: An example of a history with 10 final state microjects in a QCD event. The star vertex represents the hard interaction, the square vertices are initial state radiation and the circular vertices are timelike QCD splittings. The image is taken from [5].

There are four general types of vertices. The three final state QCD vertices (H_{ggg}, $H_{gq\bar{q}}$, H_{qqg}) and the heavy resonance decay vertex are direct approximations to the Feynman rules. The initial state radiation (ISR) vertex, which also accounts for UE, and the vertex that produces the hard parton or boosted heavy particle are modelled to fit data/Monte Carlo [5]. The first step to defining a history is to separate the radiation in the fat jet directly linked to the hard vertex from the ISR. Before a history probability calculation proceeds according to the rules, there is a condition on the maximum fraction of the radiation that can be attributed to the ISR. Let $\boldsymbol{k}_{T,I}$ be the total transverse momentum of the microjets assigned to the ISR. Then a history will be discarded if $\boldsymbol{k}_{T,I}^2 > Q^2/4$, where $Q^2 = p_{T,\,\text{fat jet}}^2 + m_{\text{fat jet}}^2$. This hard vertex approximation is only used when Shower deconstruction is applied on a single jets. A more universal approach that circumvents the need of an approximate weight for the hard matrix element is Event Deconstruction [19], where the MEM weight is supplemented by a shower deconstruction weight for each jet. Moreover, in the implementation of shower deconstruction for quark-gluon tagging in this chapter, the hard vertex cancels between the numerator and denominator.

The ISR and UE are grouped into a single type of vertex. The underlying assumption is that the radiation off the initial state is soft or collinear to the initial

state parton. In addition there is a factor that accounts for the change in parton distribution function when the emitted particle is not soft [76]. These vertices also do not contribute in this implementation.

In the final state splittings there is a complication when the emitted particle is a gluon. In particular when it is soft, the gluon is emitted from a dipole according to Eq. (1.3.18). The angular distribution of the dipole matrix element squared is

$$H_{dip}(p,p',k) \propto \frac{2p\cdot p'}{p\cdot k\, p'\cdot k} = \frac{2}{k_T^2}\frac{\theta^2_{pp'}}{\theta^2_{pk}\theta^2_{p'k}}\ . \tag{2.1.10}$$

Here we use the approximation to the invariant mass squared of two massless and nearly collinear four-vectors $2p_1\cdot p_2 \approx p_{T1}p_{T2}\theta^2_{12}$, where $\theta^2_{12} = (y_1-y_2)^2+(\phi_1-\phi_2)^2$ is the distance between the two particles in $y-\phi$. What is evident is that the dipole matrix element squared diverges when the emitted third particle k is collinear to either p or p'. For the purposes of shower deconstruction though, the emitted particle has to be associated with a single parent in order to define a history. Therefore, the amplitude is partitioned into two pieces - each corresponding to a history, in which the gluon is emitted in association with one of the particles that form the dipole. This is achieved by choosing a function $A(p,p',k)$, such that $A(p,p',k)+A(p',p,k)=1$ and, when applied to the dipole, $H_{\rm dip}(p,p',k) = H_{\rm dip}(p,p',k)A(p,p',k) + H_{\rm dip}(p,p',k)A(p',p,k)$, each term contains the collinear enhancement associated with only one of p and p'. When evaluating a history with a vertex, that involves splitting k from p, we use the first term. In an alternative history, with k splitting off from p', we use the second term. The partition function is adapted from Eq. (7.12) in [78] in the collinear approximation,

$$A(p,p',k) = \frac{\theta^2_{p'k}}{\theta^2_{pk}+\theta^2_{p'k}}\ , \quad A(p',p,k) = \frac{\theta^2_{pk}}{\theta^2_{pk}+\theta^2_{p'k}}\ . \tag{2.1.11}$$

With this in mind, the full splitting probability of a parton J into a harder (h) and softer (s) daughter partons, accounting for purely collinear as well as soft and collinear contributions, is

$$H_i = \frac{8\pi\alpha_s}{\mu_J^2}\ P_i(z)\ g_\theta\ . \tag{2.1.12}$$

The index i stands for the type of splitting (ggg, $gq\bar{q}$, qqg). The first factor contains the dependence on the virtuality of the splitting, while the second is the appropriate

(unregulated) A-P splitting function. Here z is the fraction of the harder daughter (h) transverse momentum from the parent transverse momentum $z = k_h/k_J$, $1-z = k_s/k_J$. The last factor is a non-singular angular term, remnant from the partitioning of the dipole, that depends on the distance, in $y - \phi$, of the hard particle h and the colour partner c, as well as their respective distances from the soft particle s. For ggg and qqg splittings, we have

$$g_\theta = \frac{\theta^2_{hc}}{\theta^2_{sc} + \theta^2_{sh}} \ . \tag{2.1.13}$$

This function is close to 1 when the soft particle is collinear with the mother parton and is small when the angle is larger than the separation of the particles that form the dipole. Therefore, it can be replaced by a Heaviside step function. In the case of $gq\bar{q}$ there is no partitioning because the only singularity is the collinear one, so the soft wide-angle contribution is negligible. Thus, the angular function is trivial $g_\theta = 1$. There is an explicit ordering requirement for consecutive splittings. A history will be given a non-zero weight only if at each vertex the particle that decays (J) and the particle that it originated from (K) have the following relation $\mu^2_J < 0.5\mu^2_K\, k_J/k_K$.

In order to calculate the Sudakov factors between splittings, one needs to integrate the Sudakov exponent $\mathcal{S}_j$, which is the sum of all possible ways for the particle to split. There is a different Sudakov factor associated with a quark or a gluon line. In the case of a quark, the only splitting process is $q \to qg$. Therefore, the Sudakov exponent is the negative of the integral of H_{qqg}. The integration limits on the splitting scale are set by the virtuality of the branch μ^2_J and the ordering condition, $0.5\ \mu^2_K k_J/k_K$. The z integral is performed according to the limit set by the dipole angle through the condition $\mu^2_J/k^2_J \approx z(1-z)\theta$. The gluon Sudakov exponent contains terms from the integration of H_{ggg} and $H_{gq\bar{q}}$. Therefore the two Sudakov factors are

$$\exp\{-\mathcal{S}_{ggg}\Theta(\mathcal{S}_{ggg}) - n_f\mathcal{S}_{gq\bar{q}}\}\ ;\ \ \exp\{-\mathcal{S}_{qqg}\Theta(\mathcal{S}_{qqg})\}. \tag{2.1.14}$$

The three exponents are:

$$\begin{aligned}
\mathcal{S}_{qqg} &= \frac{C_F}{\pi b_0^2}\left\{\log\frac{\alpha_s(\mu_J^2)}{\alpha_s(k_J\mu_K^2/(2k_K))}\left[\frac{1}{\alpha_s(\theta_c^2 k_J^2)}-\frac{3b_0}{4}\right]\right. \\
&\qquad\left.+\frac{1}{\alpha_s(\mu_J^2)}-\frac{1}{\alpha_s(k_J\mu_K^2/(2k_K))}\right\}; \\
\mathcal{S}_{ggg} &= \frac{C_A}{\pi b_0^2}\left\{\log\frac{\alpha_s(\mu_J^2)}{\alpha_s(k_J\mu_K^2/(2k_K))}\left[\frac{1}{\alpha_s(\theta_{c1}\theta_{c2}k_J^2)}-\frac{11b_0}{12}\right]\right. \\
&\qquad\left.+\frac{1}{\alpha_s(\mu_J^2)}-\frac{1}{\alpha_s(k_J\mu_K^2/(2k_K))}\right\}; \\
\mathcal{S}_{gq\bar{q}} &= \frac{T_R}{3\pi b_0}\left\{\log\frac{\alpha_s(\mu_J^2)}{\alpha_s(k_J\mu_K^2/(2k_K))}\right\}.
\end{aligned} \tag{2.1.15}$$

The Heaviside functions are there to remove unphysical contributions when the exponent, due to the approximations, turns negative and the Sudakov factor may explode.

The initial state parton should also be given a separate Sudakov factor because the vertex by which it decays is different from a gluon or a quark. However, because the initial scale and hard scales are fixed and there is no virtuality ordering condition explicitly imposed on the initial state, the product of ISR Sudakov factors is the same for each history. Moreover, it is the same over the signal and background models, therefore it cancels in the actual variable χ.

Finally, the heavy resonance probability to decay is modelled simply as a rectangular function of the virtuality of the decay, to account for the detector resolution as well as the microjets' ability to accurately match the momentum of the hard partons. It is normalised to 1 so that the resonance always decays,

$$P_H \equiv H\exp\{-S\} = 4\pi^2\frac{\Theta(|m_{bb}-m_H|<\Delta m_H)}{m_H\Delta m_H}. \tag{2.1.16}$$

The final ingredient in the history weight formula are the b-tagging weights $B(f_k,t_k)$. We already know the tag of the microjet either from MC or using the experimental multivariate b-tagging. Each history will produce a flavour for the final branch. Then $B(f_k,t_k)$ is the probability to get the tag given the flavour $P(t_k|f_k)$. Thus, histories with large weight from the kinematic matching may get suppressed if the flavours they assign to the final branches do not match the observation.

2.2 Analysis setup

Defining what we mean by a quark or gluon jet can be ambiguous [64, 65]. On one hand we would like to associate it with fixed order matrix element final state partons because that would provide an easy and intuitive way in the framework of Feynman diagrams to distinguish between different events. Under such an assumption, the evolution of each parton is independent of the rest of the event; therefore, a universal tagger can be defined, much like a tagger for a boosted Higgs or a top. Unlike the latter heavy particles, whose main distinguishing attribute is a decay at a specific and event-independent scale, the shapes of the jets originating from quarks or gluons are much more susceptible to long-range interactions with other parts of the event both from the initial and final states. These are related to exchange of soft and wide-angled gluons between colour connected particles. Therefore, it is quite possible that the difference in the radiation pattern of a quark between events with different colour structure is comparable to the difference in evolution between a quark and a gluon [79,80]. In order to check if an event-independent quark-gluon tagger is viable, we trained the performance of existing jet shapes and our shower deconstruction implementation on the leading jet of two types of events. The first type is a single jet with an associated Z that decays to neutrinos. The reason for choosing the Z decay channel is to facilitate the isolation of the leading jet because at this point we focus on the jet itself as opposed to any realistic event selection. Therefore the quark and gluon jets are extracted respectively from $qg \to qZ(\nu\bar{\nu})$ and $q\bar{q} \to gZ(\nu\bar{\nu})$ events. The other type of event that we use to extract quark and gluon jets is of a purely QCD type: $qq/gg \to qq$ and $q\bar{q}/gg \to gg$ respectively for the quark and gluon jet sample. Naturally in all four types of events, the flavour origin of the leading jet is unambiguous, but the colour connection patters vary. We generate the events using Pythia 8 [53] for both the matrix element calculation and the subsequent parton shower and hadronisation. Besides the event types, we also investigate the effect of the overall jet boost on the tagger sensitivity by considering jets with a lower $p_{\mathrm{T}0}$ cut of 200, 400, 600, and 1000 GeV separately.

The choice of events makes it straightforward to assign a MC flavour label to the jets. We cluster the visible final state particles into small radius Cambridge/Aachen

jets $R = 0.1$, $p_T > 1$ GeV in order to crudely approximate the angular and energy resolution of the ATLAS [81] and CMS detectors [82, 83]. Now we use these topocluster-like objects and the original final state visible particles (referred to from now on as *hadrons* even though some are leptons) as two separate sets of seeds. We perform the analysis on each set in an attempt to quantify the effect of the experimental resolution. We cluster a set of seeds into jets. The radius of the jets is another parameter that might change the radiation pattern distribution and consequently the tagging performance. Therefore, we compare the training analysis with $R_{fj} = 0.4$ and $R_{fj} = 0.8$. The event selection for Z + jet (dijet) proceeds by requiring at least one (two) Cambridge/Aachen jets to be reconstructed with a set of parameters $p_{\mathrm{T}\,0}$ and R_{fj} and rapidity $|y_{fj}| < 5$. When this condition is met, we apply energy correlation and shower deconstruction variables to the leading jet - the "fat jet". The last piece of parameterisation is the definition of the microjets in the shower deconstruction method. Obviously the result will be sensitive on how we define the fixed end-point that each shower history must reach. However, initial testing of multiple microjet definitions showed that for the purpose of quark versus gluon tagging we can use an alternative definition of shower deconstruction, where the function $\chi(\{p,t\}_N)$ is built from the four-momentum of the "fat-jet" itself $\chi(\{p,t\}_N) = \chi(p_{fj})$. For small-cone fat jets, $R_{fj} = 0.4$, this variable performs better or comparably to the full method. Choosing a wider fat jet radius reduces the performance of this new variable, therefore we revert back to the original picture of shower deconstruction and define $R_{mj} = 0.1$, $p_{\mathrm{T}mj}^{\mathrm{min}} > 5$ GeV microjets.

2.3 Observables

2.3.1 Shower Deconstruction

It was already mentioned in the previous section that a simpler instalment of χ can yield good results for quark and gluon tagging. Let us see into a little more detail what triggered the change. In the full implementation of shower deconstruction, the seeds of the fat jet are grouped into microjets using the inclusive k_T algorithm with R_{mj} and $p_{\mathrm{T}mj}^{\mathrm{min}}$. The probability for the final microjet configuration is calculated

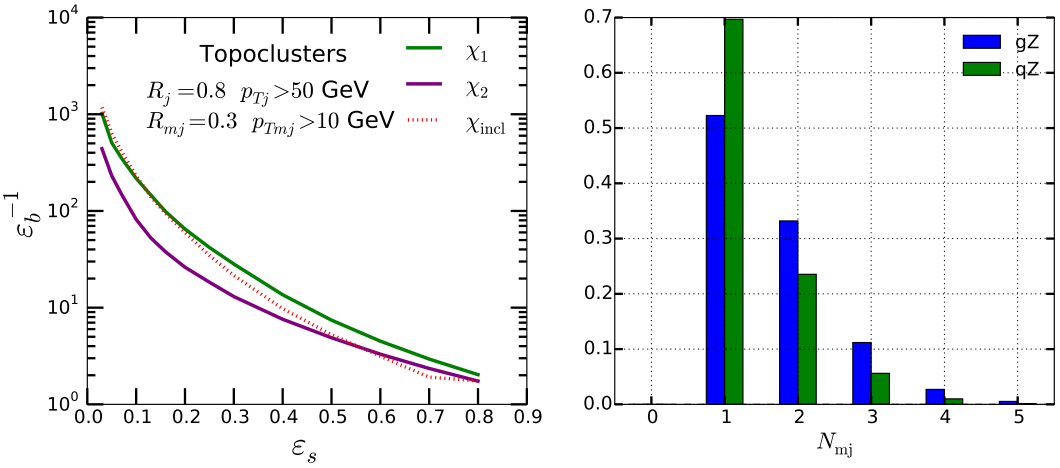


Figure 2.4: Left: quark (sig) vs gluon (bkg) ROC curves for χ with exactly one or exactly two microjets. Right: microjet multiplicity distribution.

from the sum of the probabilities of each history that leads from the momentum of the fat jet to the microjet configuration, under the assumption of a quark or a gluon. It is a well established fact that gluons radiate more as they carry a larger colour charge [24], so we expect to find that the microjet multiplicity distribution is different for qZ and gZ events. In the right plot of Fig. 2.4 we display the corresponding distributions for qZ (green) and gZ (blue). It is evident that the quark jet results in a single microjet more often than a gluon. This itself could be a discriminating feature. However, if we look in the left plot of Fig. 2.4 at normalised exclusive 1-microjet and 2-microjet samples and apply the shower deconstruction method to calculate χ_1 and χ_2, we see that the separation is much better for the 1-microjet sample. This observation led us to compare the sensitivity of χ calculated from the four-momentum of the "fat jet" to the ordinary microjet state and to find, surprisingly, that despite its simplicity, the new implementation is often a better discriminant. This behaviour differs deeply with previous implementations of the shower deconstruction method for tagging Higgs [5] and top quarks [18] from ordinary QCD jets, which relies strongly on the microjets capturing the underlying decay process.

Since this new implementation works well for quark and gluon tagging, but seems to contradict the underlying principle and general intuition behind the shower

deconstruction method, namely that finer graining of the radiation provides more useful information, it is interesting to investigate exactly what shower deconstruction measures when we use simply the four-momentum of the "fat jet". The formula for χ for just one microjet is simply a ratio of Sudakov factors:

$$\chi = \frac{P(\{p\}_m|q)}{P(\{p\}_m|g)} = \frac{e^{-\mathcal{S}_q}}{e^{-\mathcal{S}_g}} = e^{-\left(\mathcal{S}_{\mathrm{qqg}}\Theta(\mathcal{S}_{\mathrm{qqg}}>0)-\mathcal{S}_{\mathrm{ggg}}\Theta(\mathcal{S}_{\mathrm{ggg}}>0)-n_f\mathcal{S}_{\mathrm{gqq}}\right)} , \tag{2.3.17}$$

where

$$\mathcal{S}_{\mathrm{qqg}} = \frac{C_{\mathrm{F}}}{\pi b_0^2}\left\{\log\left(\frac{\alpha_{\mathrm{S}}(\mu_J^2)}{\alpha_{\mathrm{S}}(k_J^2)}\right)\left[\frac{1}{\alpha_{\mathrm{S}}(R_{\mathrm{fj}}^2 k_J^2)} - \frac{3b_0}{4}\right] + \frac{1}{\alpha_{\mathrm{S}}(\mu_J^2)} - \frac{1}{\alpha_{\mathrm{S}}(k_J^2)}\right\} ,$$

$$\mathcal{S}_{\mathrm{ggg}} = \frac{C_{\mathrm{A}}}{\pi b_0^2}\left\{\log\left(\frac{\alpha_{\mathrm{S}}(\mu_J^2)}{\alpha_{\mathrm{S}}(k_J^2)}\right)\left[\frac{1}{\alpha_{\mathrm{S}}(R_{\mathrm{fj}}^2 k_J^2)} - \frac{11b_0}{12}\right] + \frac{1}{\alpha_{\mathrm{S}}(\mu_J^2)} - \frac{1}{\alpha_{\mathrm{S}}(k_J^2)}\right\} , \tag{2.3.18}$$

$$\mathcal{S}_{\mathrm{gqq}} = \frac{T_{\mathrm{R}}}{3\pi b_0}\log\left(\frac{\alpha_{\mathrm{S}}(\mu_J^2)}{\alpha_{\mathrm{S}}(k_J^2)}\right) .$$

Here μ_J is the jet mass and k_J is the jet transverse momentum.

When we evaluate the shower deconstruction variable from the "fat jet" momentum without microjets, we see that $\chi = \chi(\mu_J, k_J)$ is only a function of the jet mass μ_J and transverse momentum k_J. In fact, to extract the sensitivity, we use the natural logarithm $\log\chi$. As already described in Sec. 2.1.2, this function is constructed to be an approximation to the log likelihood ratio between the probabilities of signal and background initiating partons to produce the reconstructed final state configuration. In general the probabilities are complicated functions of many variables, e.g. Eq. (2.1.6), but in this concrete case we have

$$\log L(q,g) = \log P_{\mathrm{MC}}(\mu_J^2, k_J^2|q) - \log P_{\mathrm{MC}}(\mu_J^2, k_J^2|g) .$$

We know from the Neyman-Pearson lemma that in the two dimensional space of the variables μ_J^2 and k_J^2, a cut on this function allows for the best separation between quark and gluon initiated jets under the condition that the Monte Carlo simulator is a good representation of nature. Unlike the generic case, here we have a very small number of variables and it is possible to numerically construct the function. Therefore, we can verify whether shower deconstruction χ is indeed a good approximation to $L(q,g)$.

In principle all we need to construct the likelihood ratio function $L(q, g)$ is to determine the probability distributions of a gluon jet and a quark jet over (μ_J^2, k_J^2). We use the normalised histograms of the leading jet in the qZ and gZ events respectively. The value of $L(q, g)$ in each bin is the ratio of the normalised histograms at that bin. Unfortunately, there are significant statistical fluctuations that distort the contours of the likelihood ratio. We attempt to ameliorate this by "spilling" part of the probability in each bin to its neighbours. To implement this probability spread, we use gaussian kernel-density estimator [84]. What this method does is to replace a delta function at the centre of each bin with coordinates (μ_{Ji}^2, k_{Ji}^2) with a 2-dimensional gaussian distribution with the same normalisation as the bin. The volume and centre of each gaussian are determined from the normalisation and coordinates of the bin it replaces, but the standard deviation is a free parameter. It controls how much of the bin is spread to the rest of the histogram, leading to a smoothing of the overall distribution. This parameter in principle should be determined by splitting the events into a training and testing samples and applying cross-validation methods to the latter. We are not so interested in the predictive power of the estimator as we are in the comparison between the shower deconstruction and likelihood ratio contours; therefore, we choose the bandwidth parameter by visual comparison with the original histogram. The results of this procedure are displayed in Fig. 2.5, where the horizontal and vertical axes represent our variables μ_J^2 and k_J^2 respectively. The bottom figure is a combination of four plots - two scatter plots and two contour plots. The red (blue) scatter plot is for the leading jet in qZ (gZ) events. The yellow lines are contours of the function $\log L(q, g)$, constructed following the steps in this paragraph. The green lines are contours of the function $\log \chi(\mu_J^2, k_J^2)$ given in Eq. (2.3.17). Our conclusion is that the latter follow the likelihood ratio contours closely enough for the shower deconstruction to be considered a good approximation to $\log L(q, g)$.

2.3.2 Energy correlation functions

Some of the state of the art techniques for discriminating quark and gluon jets are the jet shapes from the family of energy correlation functions as well as variables

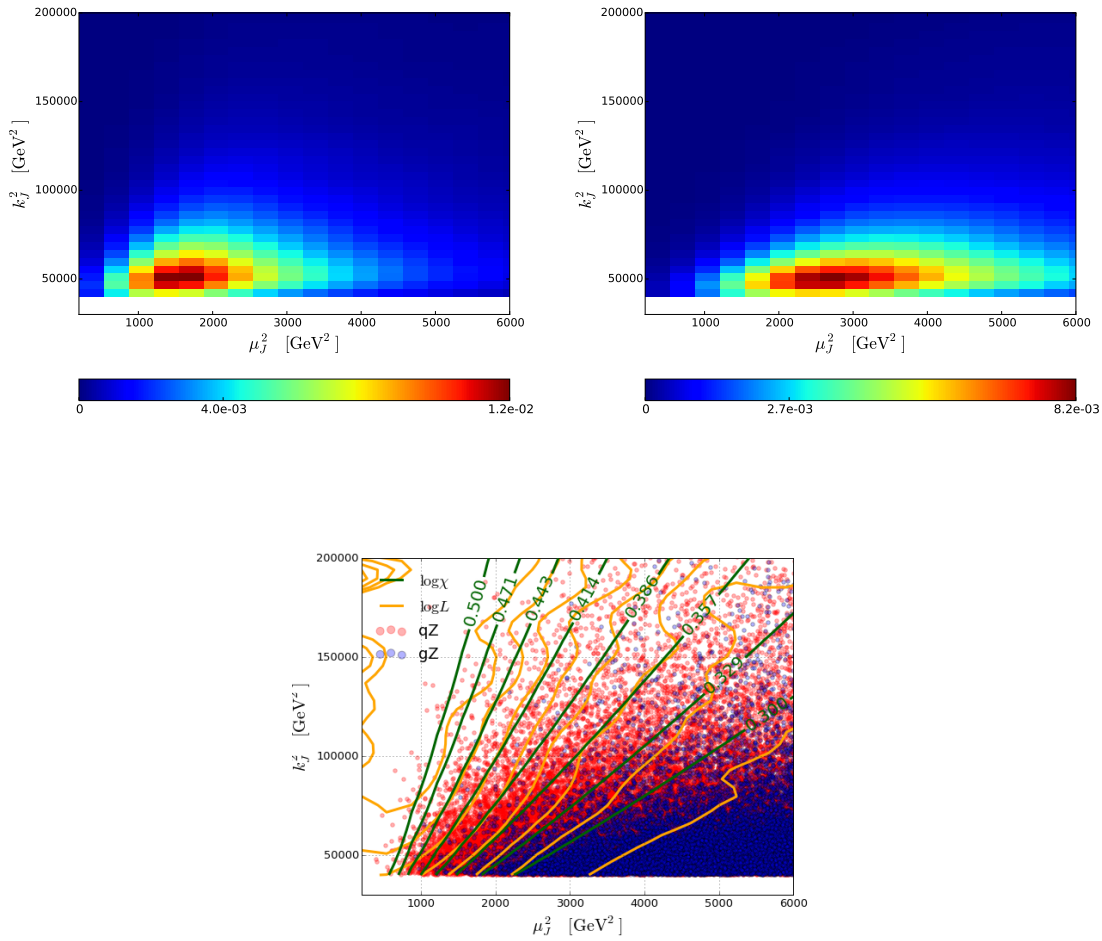


Figure 2.5: Gaussian kernel-density estimate of the leading jets' mass and transverse momentum distribution in $Z + q$ (left) and $Z + g$ (right) events. In the bottom plot we overlay a scatter plot of the two distributions, contours of the likelihood derived from the gaussian kernel-density estimator and another contour plot of the shower deconstruction variable χ.

derived from ratios of these correlations [61, 85]. The energy correlation is defined thus,

$$\begin{aligned} ECF(0,\beta) &= 1, \\ ECF(1,\beta) &= \sum_{i\in J} p_{T,i}, \\ ECF(2,\beta) &= \sum_{i<j\in J} p_{T,i}p_{T,j}\left(R_{ij}\right)^{\beta}, \end{aligned} \tag{2.3.19}$$

which generalises to

$$ECF(N,\beta) = \sum_{i_1<i_2<..<i_n\in J}\left(\prod_{a=1}^{N} p_{T,i_a}\right)\left(\prod_{b=1}^{N-1}\prod_{c=b+1}^{N} R_{i_b i_c}\right)^{\beta}.$$

The ratio and double ratio variables are the following functions,

$$\begin{aligned} r_N^{(\beta)} &= \frac{ECF(N+1,\beta)}{ECF(N,\beta)}, \\ C_N^{(\beta)} &= \frac{r_N^{(\beta)}}{r_{N-1}^{(\beta)}} = \frac{ECF(N+1,\beta)ECF(N-1,\beta)}{ECF(N,\beta)^2}. \end{aligned} \tag{2.3.20}$$

The summation in the above expressions is over the constituents i of the jet J. We explore and compare the performance of several of the jet shapes of this type (r_0, r_1, r_2, C_1, C_2, D_2), where D_2 is defined in [85], as well as other jet shapes of the N-subjettiness family [59] (τ_1, τ_2, τ_2/τ_1, τ_3/τ_2), to our results with the variable χ, defined in the previous section. The angular exponent value $\beta = 0.2$ is selected in accordance with the authors' suggestions for the application of their variables to the problem of quark and gluon tagging. Of the listed jet shapes, we find that the best performing variables are C_1, r_1, and r_2. Focussing on C_1 and r_2, we can write them out explicitly using their generic definitions in Eq. (2.3.20),

$$\begin{aligned} C_1 &= \frac{\sum_{i<j\in J} p_{T,i}p_{T,j}\left(R_{ij}\right)^{0.2}}{\sum_{i,j\in J} p_{T,i}p_{T,j}}, \\ r_2 &= \frac{\sum_{i<j<k\in J} p_{T,i}p_{T,j}p_{T,k}\left(R_{ij}R_{ik}R_{kj}\right)^{0.2}}{\sum_{i<j\in J} p_{T,i}p_{T,j}\left(R_{ij}\right)^{0.2}}. \end{aligned} \tag{2.3.21}$$

One key feature of the double ratio C_1 is that its numerator is larger if the radiation is split evenly between well separated jets than if all of it is clustered in

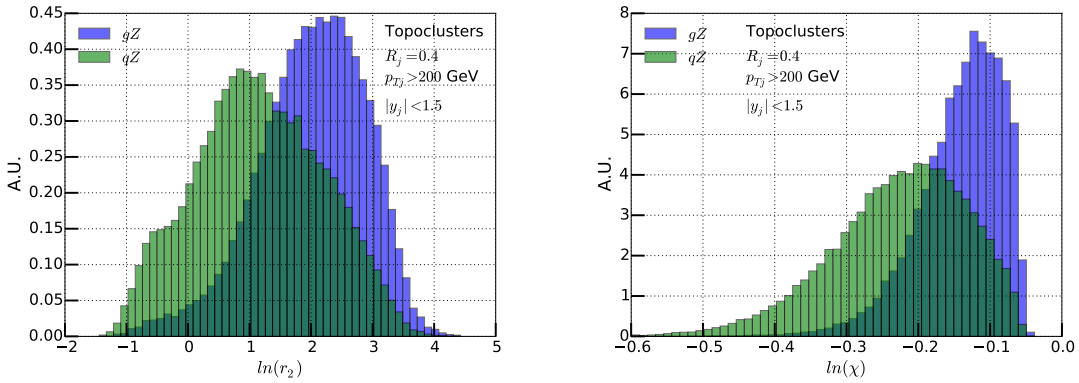


Figure 2.6: Distributions of r_2 (left) and $ln(\chi)$ (right) in $Z+$jet events. Leading jet with $|y_j| < 1.5$ reconstructed from topoclusters.

a small angular region. In the latter case for every term in the sum at least one of the three variables in the product will be small, which is not true if the radiation is split evenly. Therefore this variable changes its value drastically between 1-prong and 2-prong jets. With the same logic in mind, the other ratio r_2 is always small for 1-prong and 2-prong jets, but is not small any more once the radiation in the jet is split into three. The choice of the angular exponent $\beta = 0.2$ is suggested in Eq. (3.22) in [61]. The authors, who proposed and explored the behaviour of the C_1 variable in the Next-to-Leading Log accuracy, find a power law relation between its cumulative distribution for quark and gluon samples. Generally, the influence of a small β is to increase the power of the gluon distribution as a function of the quark distribution. This means that a cut that keeps a particular fraction of quark jets will mistag fewer gluons if β is small. The angular exponent cannot be pushed to an extremely small magnitude as the validity of the perturbation expansion would be hampered.

In Fig. 2.6 we present the distributions of r_2 and χ, applied to the leading jet in Z + jet events. The distributions are asymmetric; therefore, their sensitivity is different when used for quark tagging as opposed to gluon tagging. As a concrete example let us use shower deconstruction to improve the quark to gluon ratio and keep 20% of the signal (quarks in this case). Then we have to impose a cut $\log \chi < -0.3$, which leaves $\varepsilon_S = 0.21$ and $\varepsilon_B = 0.017$ for the singal and background

efficiencies respectively. To do the opposite, improve the gluon to quark ratio by keeping the same signal (gluons in this case), we must select jets with a $\log \chi$ bigger than a given cut that would keep exactly $\varepsilon_S = 0.21$ fraction of the gluons. In this case the quark mistag rate is three times larger at $\varepsilon_B = 0.05$. The asymmetry in the ROC curves and consequently the tagging capabilities between quark and gluon tagging is evident already in the Leading Log approximation for the C_1 variable, Eq. (3.7) in [61]. When quark tagging is performed with a cut on C_1, the background mistag rate is a function of the signal efficiency according to the power law,

$$\varepsilon_g(\varepsilon_q) = \varepsilon_q^{C_A/C_F} = \varepsilon_q^{2.25} \,. \tag{2.3.22}$$

Thus, when the quark is tagged at 50% efficiency, the gluon is mistagged as a quark at a rate $\varepsilon_g(0.5) \approx 0.21$. The reverse tagging (gluons versus quarks) requires that we perform the cut in the opposite direction of the C_1 distributions. Then the signal and background efficiencies would be given by $1-\varepsilon_g$ and $1-\varepsilon_q$ respectively, where we have kept the old definitions of ε_q and ε_g and thus the same relation between them. If we decide to make a cut that keeps 50% of the gluon jets (now the signal), then the fraction of quarks that will be labelled as gluons is $1-(1-0.5)^{\frac{1}{2.25}} \approx 0.27$. The conclusion is what we see from Monte Carlo tests, namely that the same variable can discriminate with different sensitivity depending on what particle we try to tag. This asymmetry is strongly in favour of quark tagging for all of the variables that we study, as will become evident in the following sections.

Just as with the shower deconstruction variable, a preliminary study of the energy correlation variables in relation to quark tagging allows us to limit the comparison with χ to only a couple of jet shape variables and also shows some trends in the tagging performance as the jet parameters are varied. The original energy correlation paper [61] already showed that C_1 is a good quark versus gluon discriminator. We find in accordance with it that as long as the fat jet is clustered with hadron seeds, C_1 provides the best quark tagging over large parts of the set of jet parameter choices. The comparison to r_2 is displayed in Fig. 2.7 and 2.8, where the bottom rows in particular contain the results with hadrons. Given a moderate signal efficiency, the background mistag rate from C_1 is about 60% of the rate obtained from a cut on r_2. As the cut is made more stringent, the difference disappears. All of the plots in

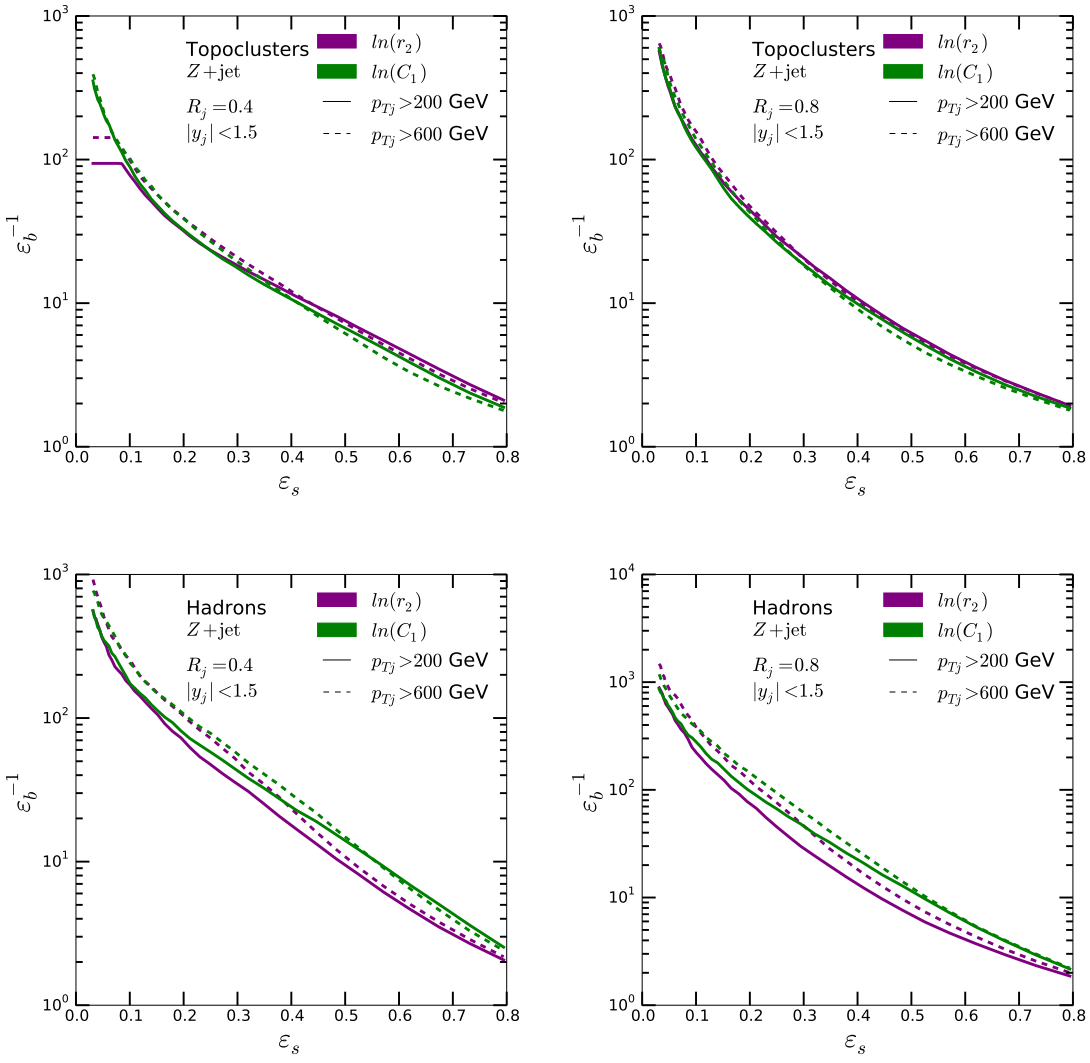


Figure 2.7: ROC plots comparing r_2 and C_1 performance at different jet p_T. The top row uses topoclusters as seeds and the bottom uses hadrons. The left (right) column uses jets with small (large) radius.

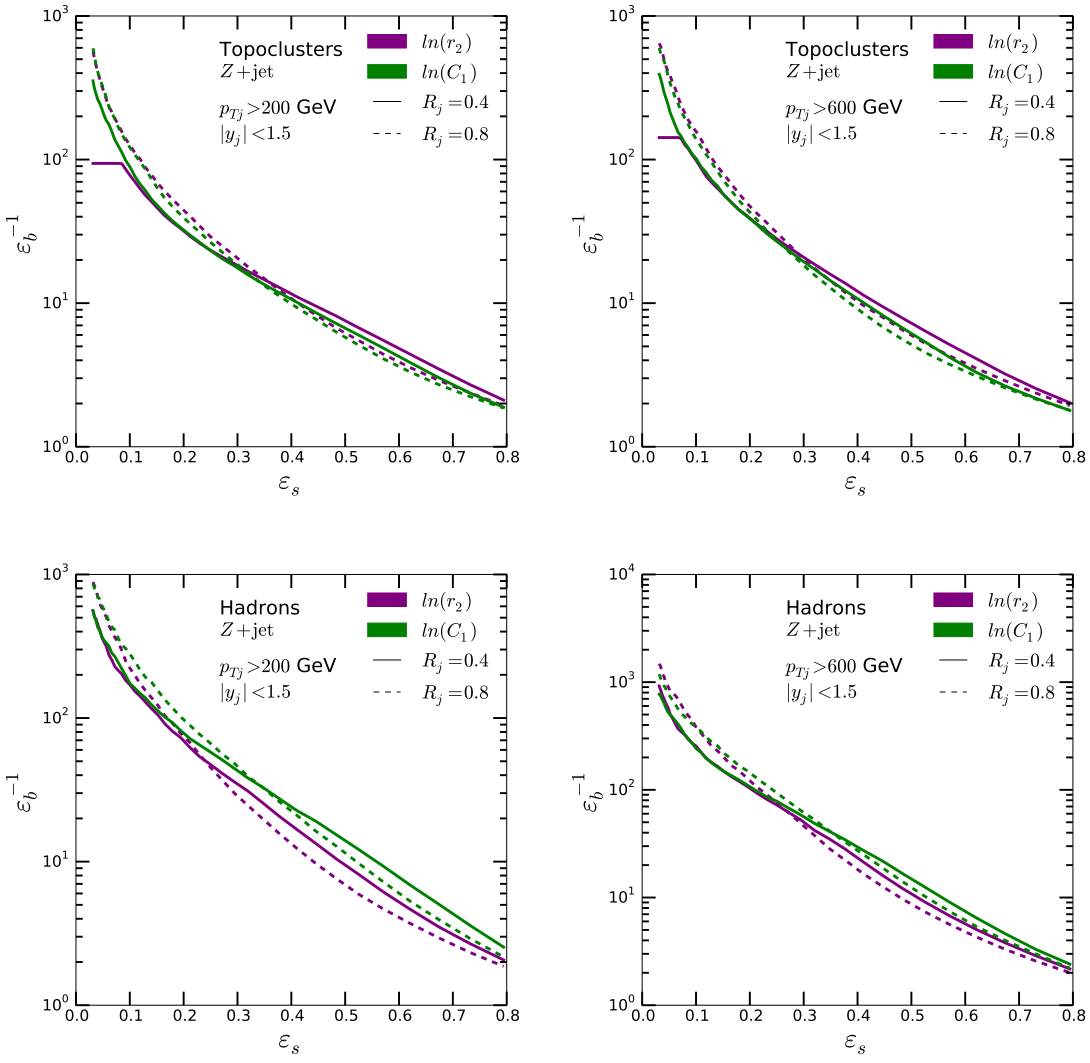


Figure 2.8: ROC plots comparing r_2 and C_1 performance at different jet radii. The top row uses Topoclusters as seeds and the bottom uses Hadrons. The left (right) column uses jets with small (large) boost.

Fig. 2.8 exhibit the same effect as the jet definition is changed to a larger radius. When that happens the energy correlation variables improve their performance at low signal efficiency at the expense of the performance with looser cuts. The effect is true for different jet p_T cut, rapidity cut and seed choice. The last trend concerns the performance as the fat jet transverse momentum limit is changed. For hadron-seed jets an increase in the p_T improves the signal to background ratio for stringent cuts. This effect is not true if the jets are reconstructed from topoclusters. For those jets the p_T limit does not alter the tagging performance of energy correlation variables.

2.4 Tagging results and uncertainties

The comparison between the performance of the listed jet shapes to the shower deconstruction variable χ can be found in the ROC curves in Fig. 2.9. These curves are built by swiping the appropriate distributions in the direction that boosts the quark to gluon ratio. For central jets, reconstructed from topoclusters, χ outperforms the other jet shape variables for all signal efficiency points. Even though Fig. 2.9 only shows a specific choice of jet parameters, the conclusion is true for a wide range as long as the jets are clustered from topocluster seeds. The shower deconstruction variable displayed here is the simplified version with the total fat jet momentum as the only input to the method. Looking at the rest of the curves, we see that no single jet shape can be distinguished as dominant. Instead, there is a tier of five variables whose gluon fake rate is within a band of $\Delta\varepsilon_b \approx 0.2$ throughout the range of quark tagging efficiency. This tier includes [r_2, r_1, C_1, τ_1, τ_2]. A closer inspection of this tier reveals that the ratio r_2 outperforms the rest, although mildly, for $\varepsilon_s > 0.3$ and is competitive at small efficiencies. Therefore, we choose to use r_2 to represent the wider family of jet shapes discussed in the previous section. This way we can focus on only two variables, r_2 and χ, and trace the difference over jet parameters, event types, and parton shower tools. When hadron seeds are concerned, we use C_1 instead.

We show the tagging performance of χ and r_2 for quark (left) and gluon (right) tagging in Fig. 2.10. Just as it was argued in Sec. 2.3.2, there is a vast discrepancy

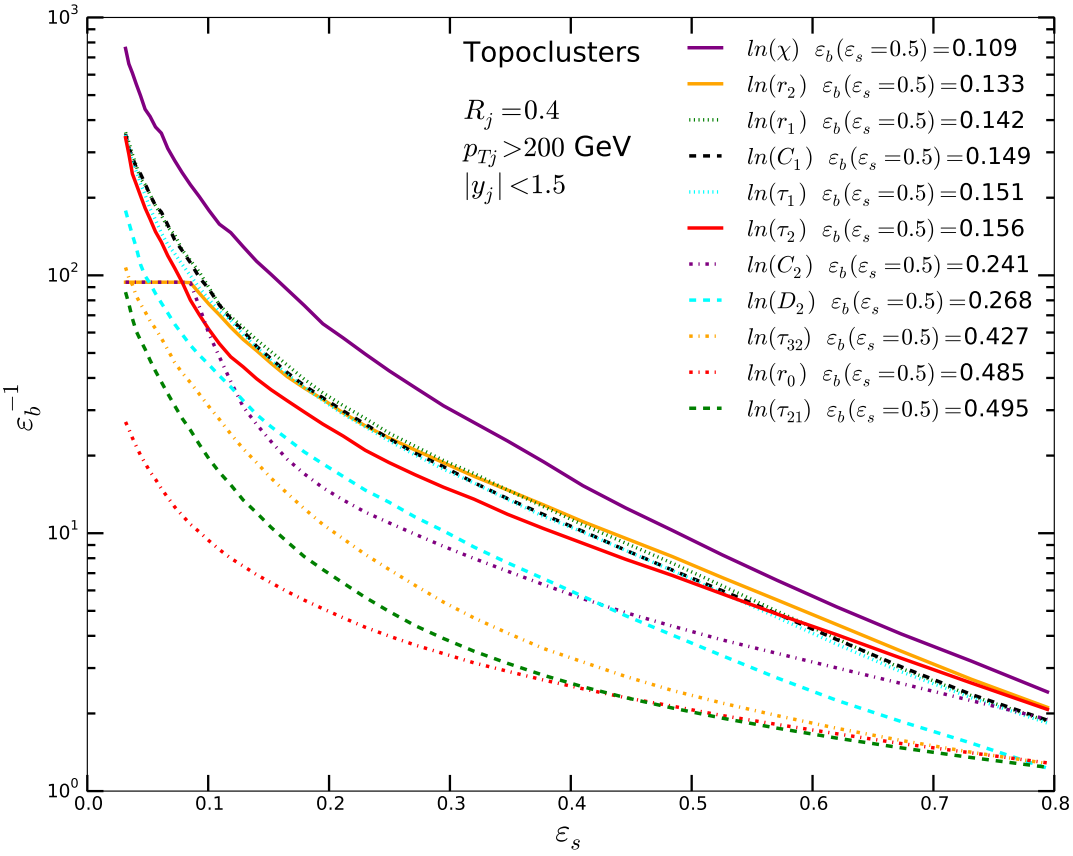


Figure 2.9: ROC curves for all distributions for quark tagging of Z + jet events. Leading jet with $|y| < 1.5$ reconstructed from topoclusters.

between the S/B ratios achieved for quark tagging and for gluon tagging. For cuts that keep only 10% of the signal, the difference between gluon rejection and quark rejection is a factor of four if the shower deconstruction variable is used. In the case of r_2 the difference is two-fold, which is smaller but still dramatic. This behaviour is evident, at least on a qualitative basis, from the probability distribution plots in Fig. 2.6, even without transforming them into ROC plots. Both χ and r_2 distributions of the quark sample drop off slower in the gluon-like end (towards larger values) than the distributions of the gluon sample in the quark-like region (smaller values). This asymmetry translates into the difference between the background rejection when performing quark tagging or gluon tagging. In particular it allows

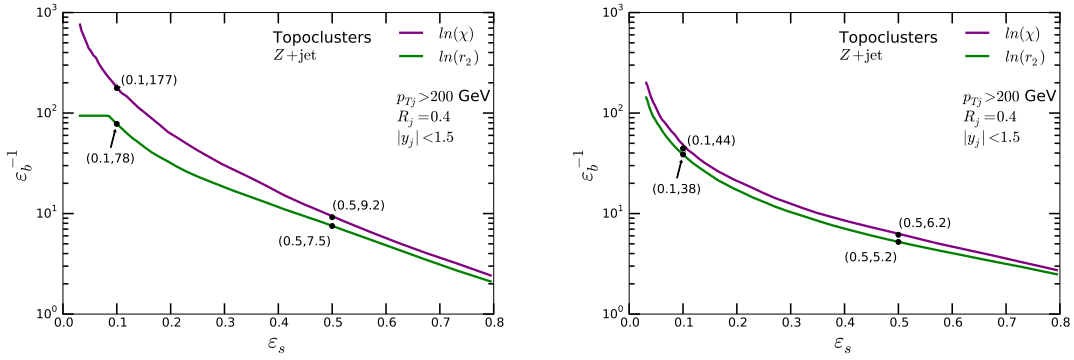


Figure 2.10: Left: ROC curves for quark tagging and gluon rejection from $Z+\text{jet}$ events. Right: ROC curves for gluon tagging and quark rejection from $Z+\text{jet}$ events. Leading jet with $|y|<1.5$ reconstructed from topoclusters.

very good gluon rejection at acceptable levels of signal retention.

We see that the single-branch χ outperforms r_2 for gluon tagging as well as quark tagging. χ provides about 20% better gluon rejection at moderate quark efficiency, which grows to a factor of two at low signal efficiency. The difference is considerably reduced as we move to gluon tagging and it practically disappears under low efficiency cuts if we replace r_2 by another, better performing, energy correlation variable in that region. A noticeably unnatural feature in the r_2 ROC curves, although in an efficiency region, which we do not explore, is the plateau at $\varepsilon_s < 0.1$. This happens because there is a bin at a large value of r_2, not shown in the distribution in Fig. 2.6, where all jets that cannot define the variable are stored. Examining the terms in the formula for r_2, we see that it only makes sense for jets with at least three constituents. This is not a problem for hadron seeds or large-radius jets, but it is quite conceivable that small $R = 0.4$ jets built from topoclusters with a large angular resolution may contain two or fewer seeds. Of course the plateau in the ROC curve is mainly an artefact of how this separate bin is incorporated into the rest of the distribution. Given our simple approach of swiping one way or another, it is rather unnaturally ordered. True optimisation of each $(\varepsilon_b, \varepsilon_s)$ point will remedy the shape of the ROC plot. Moreover, we have not attempted to optimise the angular exponent parameter in the energy correlation and N-subjettiness variables but employed the recommended value $\beta = 0.2$ for quark and

gluon tagging.

Throughout the discussion of the results so far we have only looked at χ defined on $p_T > 200$ GeV jets. The energies accessible to the LHC are an order of magnitude larger, so there can be very boosted jets either as decay products of heavy BSM particles or simply as a recoil in high p_T events. Therefore we compare the performance of χ for different jet p_T limits ranging from 200 GeV to 1 TeV. The results are presented in Fig. 2.11. We can see a very strong dependence of the background rejection on the jet p_T over the entire signal efficiency range. Coupled with the observed r_2 independence on the boost of the jet, this leads to improvements in the shower deconstruction S/B ratio at $\varepsilon_s = 0.5$ from a factor of 1.2 better than r_2 to 1.4 as we move from $p_T > 200$ GeV to $p_T > 1$ TeV jets. The effect on the performance of χ is even greater at stringent cuts, where at large boost the S/B ratio obtained with shower deconstruction is three times better than the one obtained from r_2.

So far we have considered jets that end up in the central region of the detector with $|y_j| < 1.5$. The region in the multi-purpose detectors at the LHC sensitive to jet substructure stretches to $|y| < 2.5$, so we should consider what happens when we allow for less central jets. The results with $|y_j| < 2.5$ are shown in Fig. 2.12. For low p_T jets, the performance of the energy correlation variables is not affected, while it noticeably diminishes for χ. We do not know what causes this behaviour yet, but it might be related to the fact that shower deconstruction vertices and Sudakov factors have been derived under the assumption of central jets. This means, as far as the magnitude of a microjet's momentum is concerned, that we proceed as if the beam-transverse component accounts for all of it. In other words we substitute the transverse momentum for the full momentum. Moreover, this allows us to define the splitting fraction z from the p_T ratios of the microjets involved in a vertex. The discrepancy is avoided if the jet p_T limit is increased beyond 600 GeV. Then the negative effect from opening the rapidity window goes away. The reason is that jets with such a large transverse momentum tend to have a small component in the beam direction.

Finally, we can compare the effect of widening the jet radius from $R_{\text{fj}} = 0.4$

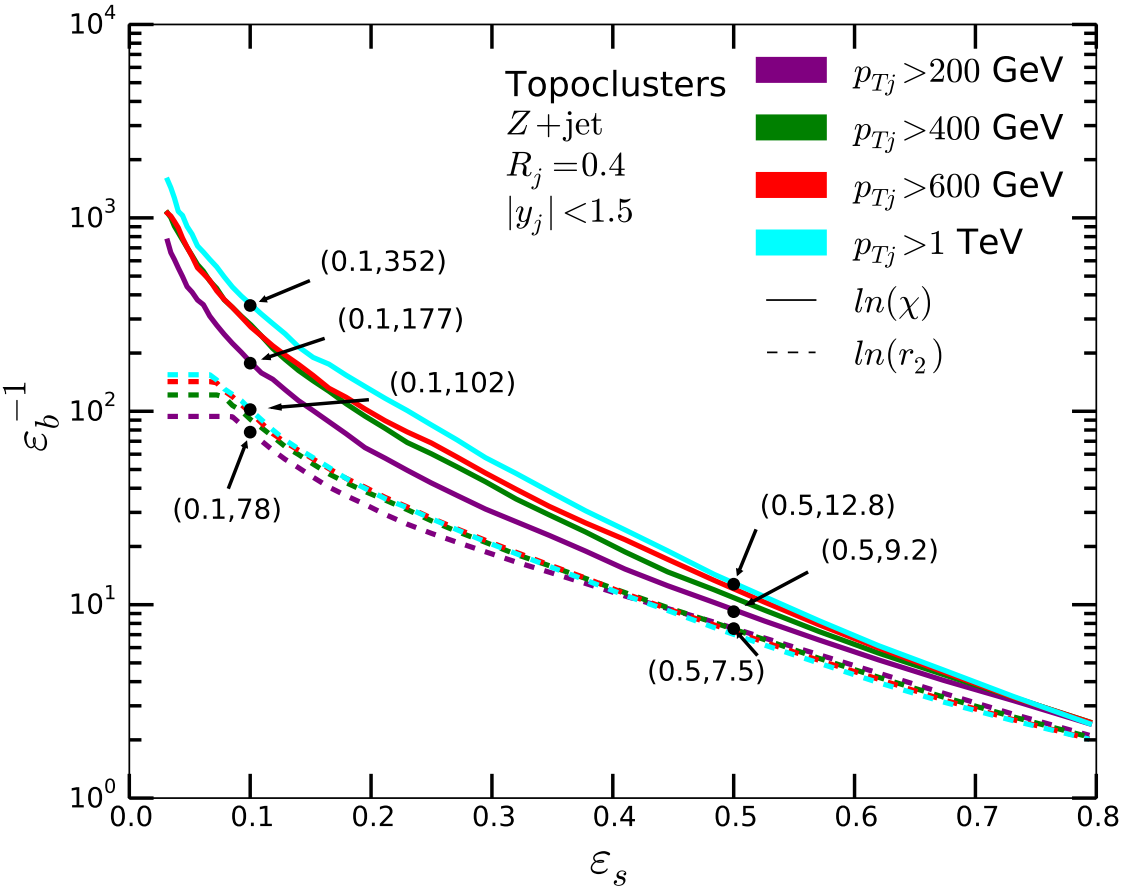


Figure 2.11: ROC curves for all p_T bins for quark tagging of Z + jet events with χ and r_2. Leading jet with $|y| < 1.5$ reconstructed from topoclusters. The solid lines correspond to $\log(\chi)$ of shower deconstruction and the dashed lines to the energy correlation function $\log(r2)$.

to $R_{\text{fj}} = 0.8$. The comparison is shown in the left plot of Fig. 2.13. We see an improvement in the r_2 gluon rejection for $\varepsilon_s < 0.4$ as the fat jet radius widens, but at the same time the performance worsens for the rest of the range. Even though increasing the jet radius seems like an even trade for r_2, the effect on χ is mostly negative apart at very low signal efficiency. Actually, we see that at moderate and large efficiencies the two variables show identical gluon rejection as long as the jet radius is large. We should consider that we have used the single-branch χ, which takes the total jet momentum as its only input. We can probe the fat jet

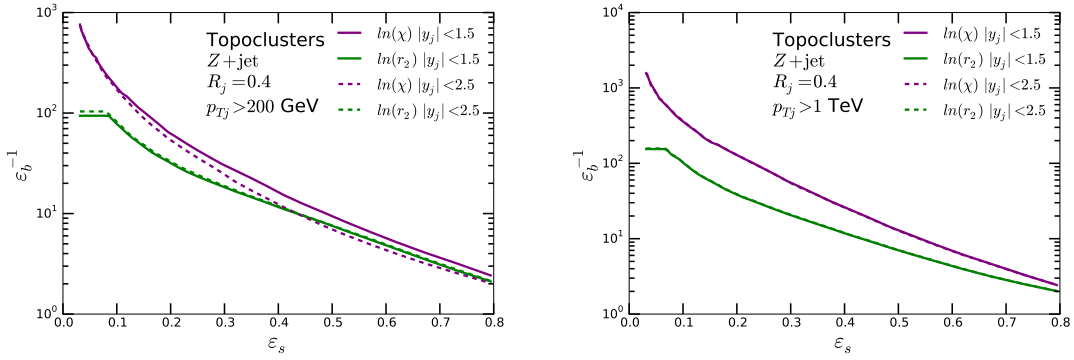


Figure 2.12: Left (Right): ROC curves for quark tagging and gluon rejection from $Z+$jet events for topocluster jets with a transverse momentum of 200 GeV (1 TeV).

substructure with smaller microjets and let the jet clustering algorithm to determine the number of microjets that go into the full shower deconstruction method. The resulting χ distributions for quark and gluon jets are not as smooth as single-branch χ. Therefore, just as in the case with r_2 at very low efficiency, we need a better algorithm to construct the ROC curves than a simple swipe in one direction. When we use multiple window cut to find an optimum background rejection for each signal efficiency point, we get the two ROC curves in the right plot of Fig. 2.13. Once again the shower deconstruction variable performs better than r_2 at any quark efficiency.

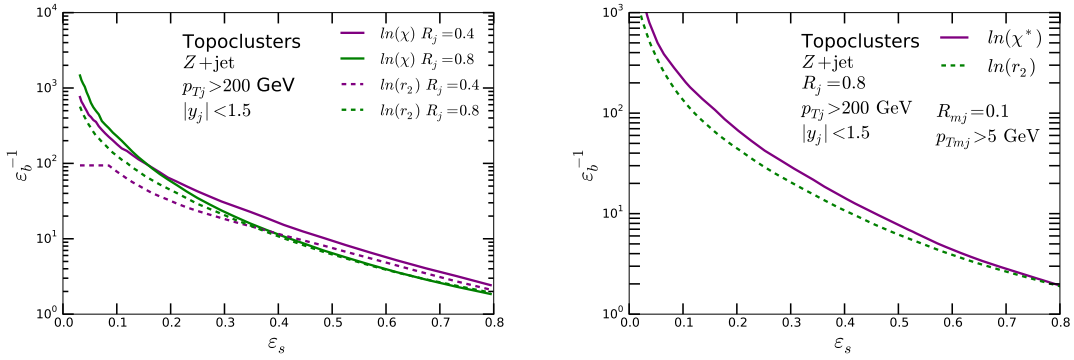


Figure 2.13: Left: ROC curves for $\log(\chi)$ and $\log(r_2)$ from $R = 0.4$ and $R = 0.8$ topocluster Cambridge-Aachen jets. Right: ROC curves from $R = 0.8$ topocluster Cambridge-Aachen jets for $\log(r_2)$ and full shower deconstruction ($\log(\chi^*)$) from $R = 0.8$ topocluster Cambridge-Aachen jets. The microjets for the true χ are Cambridge-Aachen jets with $R_{mj} = 0.1$ and $p_{Tmj} > 5$ GeV.

2.5 Results for sensitivity on underlying process and event generator

Previous studies [65] of various observables, used to separate quark from gluon jets, have identified a dependence on the collision process and the shower generator of choice. This is a serious potential source of systematic uncertainty in BSM searches or Higgs studies, which may diminish the gains from the taggers. Therefore, we investigate the dependence of the variables we compared in the previous sections on the choices of process, from which we select the jets, and the parton shower generator that we use to provide the evolution to the matrix element partons. As discussed in the analysis setup section 2.2, we check the performance of the variables for two types of events, Z + jet and dijet, and for two parton shower Monte Carlo tools, Pythia [53] and Sherpa [55].

In Fig. 2.14, we compare ROC curves for quark jet tagging in Z + jet events to that for dijet events generated with Pythia 8. The difference in the performance of any of the two variables χ and r_2 applied to the two event types is negligible to the overall difference between the variables themselves. This is some evidence for the universality of the quark/gluon taggers as their performance is unaffected when used on jets from these two underlying processes in particular. We present the results with a single jet definition for clarity, but we have confirmed that the conclusion holds for the other jet definitions discussed in the previous section.

Unfortunately, the same is not true when comparing different parton showers. In Fig. 2.15, we see that the χ ROC curve for tagging quark jets in Pythia Z + jet events is vastly different to Sherpa events. To a smaller extent the same is true for the energy correlation variable. We do not know how the shower implementation in the two generators affects the quark and gluon evolutions. It might be interesting to find where this difference comes from.

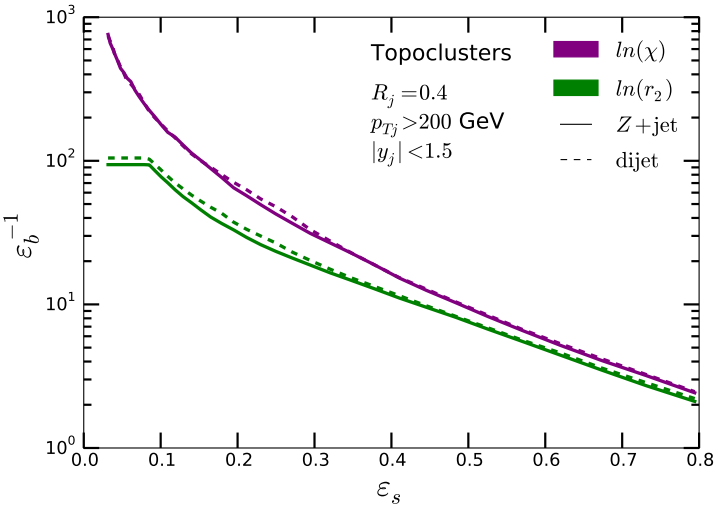


Figure 2.14: ROC curves for χ and r_2 applied to the leading jet of $Z+$jet and dijet events.

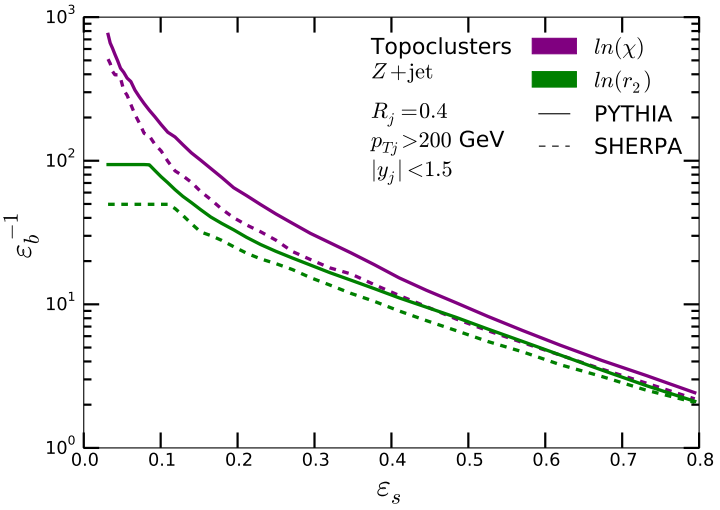


Figure 2.15: ROC curves for χ and r_2 applied to the leading jet of $Z+$jet events generated with Pythia and Sherpa.

2.6 Application of quark-gluon tagging

2.6.1 Dark matter mono-jet

One potential application of the quark and gluon tagging variable χ is in the search of dark matter candidates in the mono-jet event measurements. As a showcase we

investigate the effect of our variable on a specific simple extension to the Standard Model [86, 87]. We consider a fermion as the dark matter candidate that has no direct interaction with the Standard Model, but it is possible to produce it in pp scattering via a scalar mediator particle,

$$\mathcal{L}_{\text{scalar}} \supset -\frac{1}{2}m_{\text{MED}}^2 S^2 - g_{\text{DM}} S\,\bar{x}x - \sum_q g_{SM}^q S\,\bar{q}q - m_{\text{DM}}\bar{x}x\,. \tag{2.6.23}$$

We assume this scalar field is related to the Higgs; therefore, we expect it to couple to the SM fermions proportionally to their Yukawa coupling and we expect a very similar production phenomenology as with the Higgs boson. More concretely, the main production channel in a pp collision should be through gluon fusion into a top loop such as the bottom row in Fig. 2.1. In the case of QCD radiation that recoils with sufficient p_T against the scalar boson in the decay channel to dark matter fermions, which would pass through the detectors as invisibles, there will be a distinct mono-jet signature with uncompensated transverse momentum. To fix the parameters exactly, we choose the scalar boson to be a little heavier than the Higgs at $m_{\text{MED}} = 200$ GeV, the dark matter mass $m_{\text{DM}} = 20$ GeV and we set all couplings of the scalar boson to the fermions to be equal to their Yukawa couplings: $g_{\text{DM}} = y_{\text{DM}}$; $g_{\text{SM}}^q = y_q$. As we do not include a vertex between the heavy gauge bosons and the new scalar mediator and the top mass is too large, the main decay channel with this parameter choice is $S \to x\bar{x}$. The first column in Tab. 2.1 shows that the jet accompanying the scalar boson is equally likely to originate from a gluon or a quark.

The purely Standard Model processes that display the same detection signature are $Z(\nu\bar{\nu})+$ jet and $W(\nu l)+$ jet [67]. Given the proton pdf distributions at such a high scale, the main contribution to these processes comes from the interaction $qg \to Vq$ (Fig. 2.1) and much less often from $q\bar{q} \to Vg$. We see in Tab. 2.1 that the difference is almost an order of magnitude. What is also evident from the table is that even with such optimistic parametrisation of the dark matter extension, the SM background is still overwhelming. Coupled with the large systematic uncertainty associated with missing energy measurements [88], it is imperative to boost the signal-to-background ratio in order for a mono-jet analysis to be sensitive to

$\sigma(\text{jet} + \text{MET})$ [fb]				
13 TeV LHC				
	$p_{T,j} > 250\text{GeV}$	$\lvert y\rvert < 1.5$	$\epsilon(\chi(g,q)) \simeq 50\%$	$\epsilon(\chi(g,q)) \simeq 10\%$
$pp \to (S \to \bar{x}x)j$	190	139	46.5	8.17
$pp \to (S \to \bar{x}x)g$	96.5	78.6	36.7	6.77
$pp \to (S \to \bar{x}x)q$	93.3	60	9.27	1.14
$pp \to (Z \to \bar{\nu}\nu)j$	2830	2170	430	62.2
$pp \to (Z \to \bar{\nu}\nu)g$	334	245	122	24.6
$pp \to (Z \to \bar{\nu}\nu)q$	2460	1890	299	40.3
S/B	0.067	0.064	0.11	0.13

Table 2.1: Production cross sections for a top-philic scalar mediator of mass $m_S = 200$ GeV that decays predominantly into dark matter, see Eq. (2.6.23), and the dominant Standard Model background $Z + \text{jet}$ at $\sqrt{s} = 13$ TeV.

such extensions. According to the study by ATLAS [88], the various uncertainties associated with jet and E_T^{miss} energy resolution vary around 2%. Moreover, the accuracy of the pdf and NLO calculations of the core background processes translate to additional 3-4% background uncertainty. Even though data from the new runs will undoubtedly constrain those further, the S/B ratio needs to exceed the background uncertainty if we are to put limits on BSM models. Therefore, rejecting quark jets is vital for such studies. Even though we showed in the previous sections that boosting the gluon purity is inherently worse than quark tagging, using our variable χ we are able to almost double S/B from 0.07 to 0.13 if the collected events allow for a stringent 10% efficiency cut. With a more conservative cut that keeps 50% of the signal events, we get $S/B = 0.11$.

2.6.2 Separation of gluon- and weak boson fusion in Hjj

Quark tagging can be a very useful tool in measuring Higgs boson couplings. In particular it can be used to isolate the weak boson fusion contribution to $pp \to Hjj$ events from the gluon fusion, which has been a topic of considerable interest. An

$\sigma(pp \to Hjj)$ [fb]			
13 TeV LHC			
	$p_{T,j} > 50$ GeV, $\Delta R_{jj} > 2.0$	$\epsilon(\text{WBF}) \simeq 50\%$	$\epsilon(\text{WBF}) \simeq 10\%$
WBF $pp \to Hjj$	880	440	91
GF $pp \to Hjj$	900	180	15
GF $pp \to Hqq$	22	11	2.2
GF $pp \to Hgg$	450	61	1.8
GF $pp \to Hqg$	360	90	8
S/B	0.98	2.5	6.1

Table 2.2: LO production cross sections for gluon- and weak boson fusion of a Higgs boson with mass $m_H = 125$ GeV, separated into the respective partonic subprocesses. The two columns on the right show the results after applying a double quark tag with a combined efficiency of 50% and 10% respectively.

example of the two production modes is shown in Fig. 2.2. Therefore, we can add our quark tagger to the multitude of methods already available, such as rapidity gaps [68,72], mini-jet vetos [89,90], the matrix element method [91] and event shapes [92].

Any measurement of events involving the Higgs boson will involve multiple Higgs coupling parameters even if a particular decay channel is selected. This is because the total event count depends on the production cross section and the branching ratio, which itself depends on the total decay width as well as the coupling of the chosen decay channel. Very schematically the number of signal events from a Higgs decay channel $H \to YY$ will depend on the branching ratio and the contribution to the production cross section from each available production channel p:

$$\sigma(H) \times \text{BR}(YY) \sim \left(\sum_p g_p^2\right) \frac{g_{Hyy}^2}{\sum_{\text{modes}} g_i^2}. \tag{2.6.24}$$

We would like to make measurements dependent on as few parameters as possible; therefore, applying a cut that isolates a single production channel is an important step in studying the Higgs couplings. The weak boson fusion production channel in $pp \to Hjj$ events always produces two quark-initiated jets. In contrast the gluon

fusion almost never leads to two quarks in the final state. Therefore, a double quark tag may significantly reduce the latter.

We generate both the WBF and GF events with Sherpa. The event selection requirements are at least two Cambridge/Aachen jets with $R_{jet} = 0.4$ and loose p_T and rapidity cuts of $p_T > 50$ GeV and $|y_j| < 4.5$. We do require, however, a wide separation between the jets $\Delta R_{jj} > 2.0$. After these event selection cuts, the contribution from the two channels is almost identical. A double quark tag that leaves 50% of the WBF events already improves the purity to 70%. The last column in Tab. 2.2 shows that a more stringent cut makes the gluon fusion contribution negligible. Whether such a demanding cut can be applied will depend on the particular Higgs decay channel chosen for the study. Here we have not made such a choice and we have treated the Higgs as a stable particle.

2.7 Summary of quark and gluon tagging

We tested the performance of several established observables associated with quark and gluon tagging and compared them to a simplified implementation of the shower deconstruction method. We find that, given we use experimentally robust topocluster-like objects to construct the jets, the shower deconstruction variable χ provides better background rejection than the frontrunners in the energy correlation family r_2 and C_1. This remains true for different jet definitions as long as they fall in the central part of the detector. We have shown in Fig. 2.11 that the quark tagging capability of χ improves as the jet is more boosted in the beam-transverse direction. Even though most of the study has been performed with $R_{\text{jet}} = 0.4$ jets, the shower deconstruction method remains better performing even for fatter jets, although, in this case the full multi-microjet implementation has to be used instead of the simpler single-branch version, where the total jet momentum acts as the only microjet in the shower deconstruction framework.

We have shown that quark and gluon tagging can be useful in isolating signal events in LHC collisions in vastly different searches. Therefore, these methods can be rather universal and eventually form a procedure almost as ubiquitous as b-tagging.

However, we do not understand or control their behaviour to bring them to such status. On two occasions during our study, we noticed that the current understanding of physics at small energy scale, does not render coherent results. Specifically we see in Fig. 2.15 that the tagging efficiency from both energy correlation jet shapes and shower deconstruction is affected significantly by the choice of parton shower generator. There is a systematic difference between the quark tagging efficiency when we use Sherpa or Pythia, with the latter noticeably friendlier to gluon rejection attempts. Early on in Sec. 2.3.2, where we focussed on the energy correlation functions, we found strong dependence on the late shower evolution and hadronisation. Apparently there is some information within the hadron seeds distribution that the jet shapes can access, which is lost once the experimental resolution is taken into account. The effect can be seen in Fig. 2.7 and 2.8, where the same variables have been compared with different initial seeds. One redeeming feature, exemplified in Fig. 2.14, is that χ seems to provide a consistent background rejection when used to tag jets in different types of hard processes. This process independence is a crucial requirement for building an applicable tagger, but much more variety in the processes is necessary to claim this for sure.

Chapter 3

Collinear W tagging

With the completion of the first run of the LHC, the Higgs boson's existence was confirmed [13, 14], but any extensions to the SM have been further limited. In particular, no hints of new resonances have been seen yet, suggesting that if such exist, their masses will be at least in the TeV range. The resulting decay products will be highly boosted. In a scenario where the resonance decays to boosted top quarks, SM-initiated processes that contain boosted quark and W boson in proximity to each other, can fake a top and reduce the sensitivity to the BSM channel. The electroweak corrections are enhanced by large logarithms and a fixed order expansion in α_W is not going to provide an accurate description [93–105]. Therefore, just like in the massless QCD case, terms of type $\alpha_W \log^2(Q^2/m_W^2)$ need to be resummed to provide a dampening exponential Sudakov factor and an accurate interaction rate.

Such Sudakov factors are included in the parton shower of event generators [55, 106, 107]. This allows us to build and test techniques that look for W bosons that are collinear with boosted quarks. This type of W tagging can be useful in measuring the collinear W emission rate and compare to the predicted cross section with the appropriate phase space cuts. Moreover, finding a W in the vicinity of a jet will indicate that the jet originates from a quark. We refrain from attempting quark-gluon tagging, but we do vary the splitting function with a multiplicative factor in order to check how sensitive our analysis is to discrepancies in the measured rate.

We generate dijet events $pp \to jj$ at $\sqrt{s} = 14$ TeV with a modified version of Sherpa [55]. The matrix element is computed with Comix [108] and the partons are

showered with CSShower [109, 110], which is modified to apply an EW shower in addition to the QCD and QED shower. After that, the partons are hadronised [111] and UE [112] is also incorporated for a more realistic search.

The EW parton shower approximates the cross section for a heavy gauge boson emission by factorising the emission from the rest of the process

$$\mathrm{d}\sigma_{n+V} = \mathrm{d}\sigma_n \sum_f \sum_s^{n_{\mathrm{spec}}} \frac{\mathrm{d}t}{t}\,\mathrm{d}z\,\frac{\mathrm{d}\phi}{2\pi}\,\frac{1}{n_{\mathrm{spec}}}\,J(t,z)\,\mathcal{K}_{f(s)\to f^{(\prime)}V(s)}(t,z)\,. \tag{3.0.1}$$

The sums run over the fermions in the final state signifying the emitting parton (f) and the possible n_{spec} spectators (s). The emission probability is a result of the scale of the splitting t, the splitting fraction z and the azimuthal angle ϕ. There is also a Jacobian [110] for the transformation of the one-particle phase space element from $d^3p \to dt\,dz\,d\phi$. The exact choice of t and z, and consequently $J(t,z)$, varies between initial and final state participants in the splitting. Finally the dynamics of the emission are collected in the splitting function $\mathcal{K}$ [113]

$$\mathcal{K}_{f(s)\to f'V(s)}(t,z) = \frac{\alpha}{2\pi}\left[f_V\,c_\perp^V\,\tilde{\mathrm{V}}^{\mathrm{CDST}}_{f(s)\to f'b(s)}(t,z) + f_h\,c_L^V\,\tfrac{1}{2}\,(1-z)\right]\,. \tag{3.0.2}$$

This is schematically true for both $V = W$, Z bosons. The functions $\tilde{V}^{\mathrm{CDST}}_{fs\to f'bs}$ are derived in [114, 115]. We checked that the contribution from transversely polarised W bosons supersedes the longitudinal as well as all of the Z boson polarisations. Therefore, we focus exclusively on transverse Ws by setting $f_h = f_Z = 0$. Then the only parameter left is $c_\perp^W$, which is the combination of coupling factors associated with the W, $c_\perp^W = s_{\mathrm{eff}}\,\frac{1}{2s_W^2}\,|V_{ff'}|^2$. $s_{\mathrm{eff}} = 1/2$ accounts for the fact that the dijets are unpolarised but the W couples to the left-handed quarks only. The remainder of the chapter focusses on ways to improve the sensitivity to the emission rate factor $f_W \equiv f$, which is $f = 1$ in the SM, but we generate events with different values.

3.1 W reconstruction in dijet events

We cluster the final state radiation of each simulated event into objects that mimic the experimentally detectable ones. In particular we identify an electron or a muon as an isolated lepton when it has $p_{Tl} > 25$ GeV, it is within pseudorapidity $|\eta_l| <$

2.5 and crucially the hadronic radiation within a cone of radius $R = 0.2$ around the direction of the lepton in $\phi - \eta$ space contributes less than 10% of the lepton transverse energy. All leptons that comply with the isolation criteria are removed from the rest of the visible particles. Of those we keep the particles with transverse momentum $p_T > 0.5$ GeV and absolute pseudorapidity $|\eta| < 5.0$ and cluster them into tiles of size $\Delta\eta \times \Delta\phi = 0.1$ to account for the energy and angular resolution in an experimental setting. We use the cells as seeds to reconstruct jets. We take into account the trigger efficiency by initially requesting that an event has at least one anti-k_T jet with radius parameter $R_{jet} = 1.5$ and $p_T > 200$ GeV. If this requirement is satisfied, we separate the analysis into two mutually orthogonal parts depending on the number of isolated leptons. If we cannot reconstruct any such leptons, we perform an analysis tailored to detect a hadronic W decay. Alternatively, if we find exactly one isolated lepton among the event remnants, we attempt a leptonic W tagging. Both independent regions are further subdivided according to a minimum fat jet p_T requirement. We accept events with at least two fat jets and bin them according to a minimum jet transverse momentum limit $p_T > 500,\ 750,\ 1000$ GeV. Therefore, an emitted W boson will be boosted more frequently as we move up from the low to the high p_T bin. Note that unlike the event binning according to the number of isolated leptons, these p_T bins are not independent. All the events in a higher-p_T bin are also included in all lower p_T bins.

3.1.1 Hadronic analysis

Two ways, in which to approach the hadronic W boson identification, are to look for the signature mass scale of the heavy particle by grouping its remnants appropriately and to study the general energy distribution among those remnants with jet shapes. Even though these approaches carry some redundant information, there is still information to be gained by combining the methods into a single analysis. Therefore, while the mass search techniques yield better results than cuts on jet shapes when applied independently, we find that a consecutive application of both is the best strategy.

We devise three mass search strategies, each suited to a particular kinematic

regime: a highly boosted $p_{T_W} \gg m_W$, a moderately boosted $p_{T_W} > m_W$, and slightly boosted $p_{T_W} \simeq m_W$. The first two employ sub-structure mass reconstruction methods, while the last attempts an event-wide mass reconstruction.

(A) To look for the most boosted scenario, we cluster the fat jet constituents into $R = 0.5$, $p_{T_i} > 200$ GeV C/A subjets. The most energetic subjet will be from the emitting quark, while the second is expected to be the hadronic W. If most of the W mass is to be contained in a single subjet, then its boost must be about $p_{T_W} \geq \frac{2m_W}{R}$, so we require the fat jet to contain at least two subjets with $p_T > 200$ GeV. Then we apply the BDRS algorithm [116] to the second subjet and accept a W candidate only when the mass of the BDRS-treated subjet is within the window $m_{\text{BDRS}} \in [74, 90]$ GeV.

(B) In order to reconstruct hadronic W bosons that are not boosted enough to fit in a single subjet, we try to associate the immediate W products with even smaller-radius subjets. Therefore, for the moderately boosted case we recluster the fat jet constituents into $R = 0.3$ and $p_T > 20$ GeV C/A subjets. We will refer to this set of subjets as *microjets*. The hardest microjet is again associated with the emitting quark; therefore, we discard it. A study into the order of collinear emissions [106] reveals that it is more likely that a highly boosted quark will emit a W boson at a larger scale than a gluon. If this W boson goes on to split symmetrically into two quarks, they will usually be the seeds for the second and third microjets. We expect that the mass distribution m_{23} of the combined four-momentum of microjets two and three will show a peak structure around the W mass, so we use it as the discriminant. The mass window cut that leaves the best signal to background ratio, accounting for the mass binning limitations in an experiment, is $m_{23} \in [70, 86]$.

(C) Finally we consider a W emission without significant boost. In this case the boson may be emitted at a large radial distance from the quark. Coupled with the fact that the W decay products have more freedom to travel in a direction different from their mother boson, it is unlikely that the original quark and emitted W will form a single fat jet. Therefore, we recluster the entire event

into small $R = 0.4$, $p_T > 40$ GeV anti-k_T jets. We require at least five such jets to consider the event as a dijet + soft W candidate. Because of the original dijet event selection criteria explained in the beginning of this section, we expect that the first two most energetic jets to come from the two boosted quarks and we ignore them. We pair up the remaining jets and define the invariant masses $m_{kl}^2 = (p_k + p_l)^2$. Given the large LHC scattering scale, QCD radiation can often occur at a virtuality comparable with the W mass. Therefore, the more pairs of jets in an event we examine, the greater the chance of finding at least one with invariant mass in the proximity of the W boson and cause our method to mistag pure QCD as a W boson. In order to avoid unnecessarily biasing the QCD background we restrict the possible jet pairs. First, we only use jets three through six (or five if an event contains only five jets), $k \in [3, 6]$. Moreover, we avoid m_{34} as it is very likely that the third or the fourth hardest jet is a gluon radiation from the quark that did not emit the W boson. The only viable masses are then m_{3l} and m_{4l} where the label l refers to jets 5 and 6. We count the event as containing a W boson if one or more of the viable pairs has a mass within the range $m_{kl} \in [70, 86]$ GeV. In the case of more than one, we take the pair of jets with the smallest $\Delta m = |m_{kl} - m_W|$ to be our W candidate and label the mass variable m_{min}.

As already pointed out in the description of A, this method is increasingly more effective when the W is more boosted. Following the approximate radial separation of two-body decay products, method A can hope to find W bosons with a transverse momentum of at least $p_{T_W} \geq 300$ GeV. Fig. 3.1 shows the resulting distributions for m_{BDRS}, method A, in the three different fat jet p_T selections $p_T > 500,\ 750,\ 1000$ GeV. There is some freedom in the parametrisation of the BDRS mass. To fix this freedom we follow the original paper [116] and choose $(\mu, y_{cut}) = (0.54, 0.13)$ for subjets with $200 < p_{T_i} < 500$ GeV and $(\mu, y_{cut}) = (0.72, 0.09)$ for the rest. There is an excess of events around $m_{\text{BDRS}} = 80$ GeV, whose magnitude increases with the multiplicative factor in the splitting function f. This is expected as a higher f corresponds to more frequent EW emissions. The reconstructed W mass peak is more pronounced as we increase the fat jet p_T limit

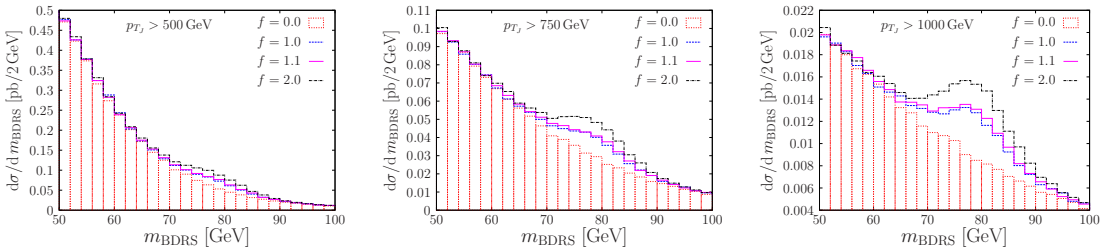


Figure 3.1: W candidate mass distribution using method A for $p_{T_J} > 500$ (left), 750 (center) and 1000 (right) GeV.

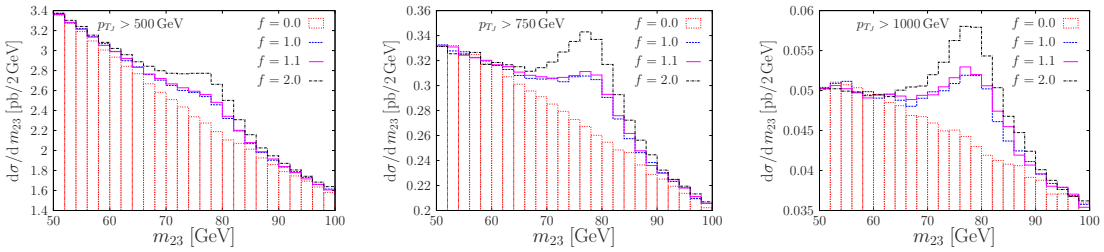


Figure 3.2: W candidate mass distribution based on microjets ι_2 and ι_3 as described in method B for $p_{T_J} > 500$ (left), 750 (center) and 1000 (right) GeV.

(moving from left to right in Fig. 3.1). This is expected because a more boosted quark has a larger possibility to emit a collinear boosted W boson, exactly the type that the BDRS method is supposed to tag.

The same information, but for the mass variable m_{23} of method B, is presented in Fig. 3.2. Just as with method A, a stronger quark boost allows for more frequent production of W bosons whose decay products are boosted enough to form the second and third microjets. Therefore, the pair is the true W more often and the peak is more pronounced as the fat jet p_T limit goes up. At first glance the mass peak for the highest p_T bin is larger with method B, but upon comparing the $y-$axis scales and starting points, we can see S/B is comparable between the two methods.

The mass distribution from the final method C can be found in Fig. 3.3. Unsurprisingly, the peak improves with larger f as the rate of EW emissions increases. On the other hand there is a contrast with the previous two methods when it comes to the quark boost effect on the sharpness and scale of the peak. Because the last method does not assume a boosted W, it does not improve performance under a

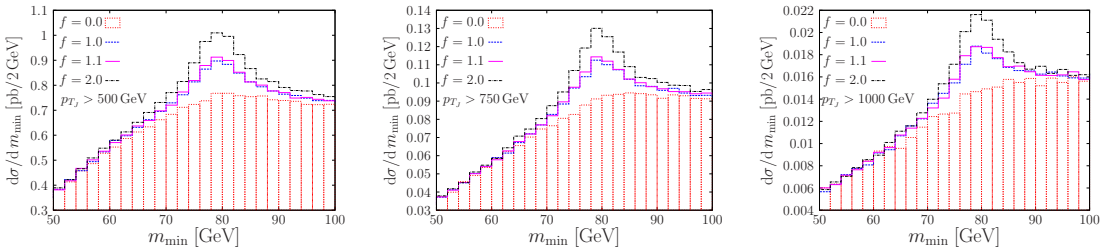


Figure 3.3: W candidate mass distribution based on method C for $p_{T_J} > 500$ (left), 750 (center) and 1000 (right) GeV.

stricter fat jet transverse momentum cut.

As discussed in the beginning of the section, combining subjet mass searches (methods A and B) with additional jet shape cuts strengthens the signal extraction. In order to link the two, we evaluate the jet shape observables only on the constituents of successfully identified, according to the method definition, hadronic W bosons. Ellipticity $\hat{t}$ (Appendix C) and the N-subjettiness ratio $\tau_{21} = \tau_2/\tau_1$ [59] provide the best additional separation when applied to the constituents of the $R = 0.5$ subjets with $m_{\text{BDRS}} \in [74, 90]$ GeV defined in the procedure of method A. We show both the ellipticity and N-subjettiness ratio distributions in Fig. 3.4. In both cases the total cross section under the curves increases with the multiplicative factor f, which is the consequence of the mass reconstruction cut acquiring more events as the rate of emission increases. The second and more important feature is that the shapes changes as well. There is a distinct shift in the peak of both distributions to lower values as the emission rate increases.

The origin of the effect is the same. The ellipticity is defined in such a way that if the radiation within the jet is clustered in one plane, or from the point of view of the jet transverse plane the transverse components lie along a single line, the observable will have a smaller value than when it is applied to a jet with isotropic energy distribution. In the perfect case scenario, a symmetric two-body decay of a massive colour singlet particle will have most of its energy within a band that stretches between the decay products. This energy profile translates into a small ellipticity value. The background that we try to discard is a high virtuality gluon mimicking the W mass. Such a particle has no fundamental scale that would

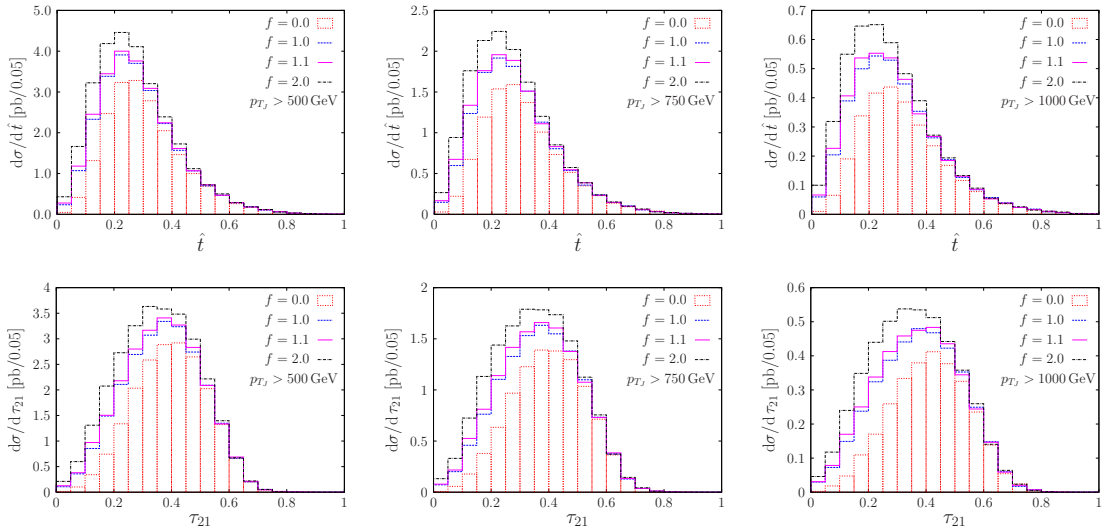


Figure 3.4: Ellipticity $\hat{t}$ (top row) and τ_{21} (bottom row) distributions calculated using constituents of W candidates identified with method A for $p_{T_J} > 500$ (left), 750 (center) and 1000 (right) GeV.

favour a symmetric split. Moreover, the gluon is colour connected to other particles; therefore, the second consecutive radiation in the gluon shower is not bound to end up between the previous two branches. Therefore, gluon jets do not have this one-dimensional profile in the jet transverse plane and are more likely to obtain a large ellipticity value. It is expected then that a sample richer in hadronic Ws (larger f) will have an ellipticity peak at smaller values than a W-depleted sample (for example $f = 0$).

The trend for a shift to lower τ_{21} as the W emission rate increases can also be explained with the fundamental mass scale in a W jet. The N-subjettiness variables are such that for any particle distribution $\tau_{N+1} \leq \tau_N$. In a 1-prong jet, a more probable QCD outcome, adding a second axis will not drastically change the distance of many hard particles to the closest axis. Therefore, the relation between 1-subjettiness and 2-subjettiness is $\tau_2 \lesssim \tau_1$. If the radiation in the jet is 2-prong, then two axes will substantially lower the distance between most particles and an axis. Therefore, $\tau_2 \ll \tau_1$. Thus, the ratio τ_{21} has a peak at small values as the fraction of W jets in the sample increases.

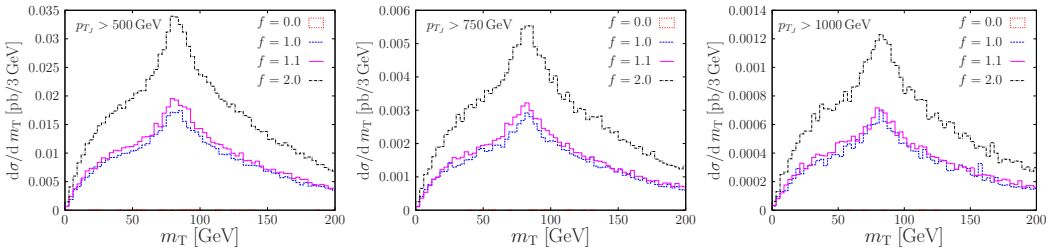


Figure 3.5: Transverse mass of the leptonic W candidate m_T for $p_{T_J} > 500$ (left), 750 (center) and 1000 (right) GeV.

3.1.2 Leptonic analysis

In order to perform the leptonic analysis instead of the hadronic, we require that the event has a single isolated lepton with $p_{Tl} > 25$ GeV and $|\eta_l| < 2.5$. Since one of the decay products in a leptonic W is a neutrino, such an event would have a signature missing transverse energy. Therefore, to proceed we also require that $\not{E}_T > 50$ GeV. Unfortunately, the component of the neutrino momentum parallel to the beam axis cannot be reconstructed from momentum conservation principles. Still, the fundamental mass scale of the W boson will show itself in a transverse mass distribution,

$$m_\text{T} = \sqrt{2E_{T_l}\not{E}_T\,(1-\cos\theta)}, \tag{3.1.3}$$

where θ is the angle between the missing energy vector and the isolated lepton. The distribution of this variable m_T has a peak structure in the vicinity of the W mass as we see in Fig. 3.5. We accept the pair of missing energy and isolated lepton as a successfully tagged W as long as its transverse mass is within the bin $m_T \in [60, 100]$ GeV. The final acceptance rates for this analysis are show in Table 3.3, where we can see that virtually no pure QCD events ($f = 0$) survive cuts. Therefore, this approach vetoes all jets with no EW emissions when the Sudakov factor has a realistic W splitting probability contribution $f \gtrsim 1.0$. When the emitting quark boost is large and the leptonic W is collinear, the proximity of the hadronic radiation from the quark evolution to the W charged lepton is going to reduce the signal efficiency of the isolation criterion. In this highly boosted regime it might be statistically beneficial to use a dynamic [117] isolation criterion.

3.2 Measuring W boson emission rates

The cross section that remains after the various cuts and selections in the analysis is summarised in the three tables of this section. A common feature in all is that each row corresponds to a different value of the multiplicative constant $f \equiv f_W$ defined in Eq. (3.0.2). Table 3.1 shows the effect of the trigger, isolated lepton and dijet minimum transverse momentum cuts. The columns $n_l = 0$ and $n_l = 1$ signify the separation of the cross section into a hadronic and leptonic bin respectively. At this point the only selection criteria that are satisfied are the trigger requirement that the event should contain at least a single $p_T > 200$ GeV jet and also the required number of isolated leptons. Furthermore, in each of the hadronic and leptonic cases we examine three different (but not independent) regions defined by a minimum p_{T_J} condition on the fat jet.

The next table is dedicated to the different versions of the hadronic analysis described in Sec. 3.1.1. For each method we keep track of the remaining cross section after a mass cut in all three p_{T_J} bins. Obviously the cross section is affected by the fat jet p_{T_J} limit. There is also an effect due to the p_T requirements on the subjets in the different methods: $p_{T_\iota} > 200$ GeV in A; $p_{T_\iota} > 40$ GeV in C; $p_{T_\iota} > 20$ GeV in B. The cross section that remains after the application of those methods increases as the p_T requirement on the subjets is relaxed. The ratio between methods B and C remains close to three for all fat jet p_{T_J} bins and multiplicative factor values. The first method does not keep a constant proportion with the other two. The ratio in the lowest $p_{T_J} > 500$ GeV bin between A and B is much smaller than the same ratio in the middle p_{T_J} bin, which in turn is yet again smaller than the ratio corresponding to the highest p_{T_J} bin. This is because the condition to have a second $p_{T_\iota} > 200$ GeV subjet with a large radius $R = 0.5$ in a fat jet that is already only $p_{T_J} \gtrsim 500$ GeV is very restrictive. If we trace back the transverse momentum of the fat jets that get a positive W identification with method A in the lowest p_{T_J} bin, then we see that half of them actually have $p_{T_J} > 750$ GeV. In contrast, the successfully tagged hadronic Ws with method C in the lowest p_{T_J} bin stem from fat jets with $p_{T_J} < 750$ GeV 90% of the time. Finally, the third table shows the numbers for the leptonic events after the transverse energy cut $\not{E}_T > 50$ GeV and the consecutive transverse mass

cut. We also keep track of the fat jet p_{T_J} limit.

	hadronic				leptonic			
f	$n_l = 0$	p_{T_J} [GeV]			$n_l = 1$	p_{T_J} [GeV]		
		500	750	1000		500	750	1000
0	2116	551.2	59.53	10.24	0.001	0.002	0.0002	3×10^{-5}
1.0	2092	539.1	57.74	9.856	23.37	3.663	0.5795	0.1286
1.1	2090	537.9	57.57	9.826	25.73	4.056	0.6341	0.1389
2.0	2070	527.5	56.00	9.481	45.71	7.081	1.117	0.2439

Table 3.1: Cross sections of the hadronic and leptonic analyses in pb. Where applicable a column has three numbers to account for different fat jet p_T cuts: $p_{T_J} > 500$ (left), 750 (middle) and 1000 (right) GeV.

	method A ($m_{\text{BDRS}} \in [74, 90]$ GeV)			method B ($m_{23} \in [70, 86]$ GeV)			method C ($m_{\text{min}} \in [70, 86]$ GeV)		
f	p_{T_J} [GeV]			p_{T_J} [GeV]			p_{T_J} [GeV]		
	500	750	1000	500	750	1000	500	750	1000
0	0.9939	0.4906	0.1447	35.87	4.228	0.6943	11.81	1.401	0.2255
1.0	1.219	0.6202	0.1923	38.83	4.698	0.7890	13.22	1.607	0.2643
1.1	1.251	0.6386	0.1977	39.11	4.741	0.8000	13.34	1.623	0.2661
2.0	1.422	0.7312	0.2286	41.43	5.085	0.8584	14.49	1.780	0.2939

Table 3.2: Cross sections after the three mass reconstruction cuts in the three different methods for the hadronic analysis in pb. Each column contains three numbers to account for different fat jet cuts: $p_{T_J} > 500$ (left), 750 (middle) and 1000 (right) GeV.

f	$\not{E}_T > 50$ GeV			$m_{\mathrm{T}} \in [60, 100]$ GeV		
	p_{T_J} [GeV]			p_{T_J} [GeV]		
	500	750	1000	500	750	1000
0	0.001	1×10^{-5}	4×10^{-7}	6×10^{-5}	5×10^{-6}	1×10^{-7}
1.0	2.062	0.3481	0.07988	0.5769	0.09271	0.02156
1.1	2.280	0.3795	0.08654	0.6402	0.1046	0.02323
2.0	4.000	0.6765	0.1531	1.108	0.1830	0.04099

Table 3.3: Cross sections after the $\not{E}_T > 50$ GeV cut and the m_{T} cut in the leptonic analysis in pb. Each column contains three numbers to account for different fat jet cuts: $p_{T_J} > 500$ (left), 750 (middle) and 1000 (right) GeV.

All of the reconstructed mass cuts in the three hadronic analyses keep enough cross section that the expected integrated luminosity in Run 2, $\int \mathcal{L}dt \approx 100\,\mathrm{fb}^{-1}$, should provide statistically sufficient number of hits. This is true even for the very boosted case of $p_{T_J} > 1$ TeV fat jets. Therefore, the sensitivity of our analysis to discrepancies in the detected and expected electroweak emissions in dijet events will be limited by systematic uncertainties and the signal-to-background ratio. Given the peak structure in the mass distributions, it is conceivable to apply a side-band analysis to avoid theoretical uncertainties in the QCD background.

We estimate the sensitivity of the different approaches we described in the previous section using binned log-likelihood ratio as the test statistic, q_{W}, in a hypothesis test performed according to the modified frequentist method [118], also known as the CL_s method (Appendix A). We calculate the median exclusion sensitivity of a hypothesis with $f \neq 1$ from the Standard Model hypothesis $f = 1$. In addition to treating each bin as a separate counting experiment, we also follow the treatment of systematic uncertainty as a nuisance parameter with a gaussian distribution. Far from being an exhaustive treatment of potential sources of systematic error, this is a quick guide to what level of control over the systematic effects is needed for exclusion of different f values.

Before moving to exclude $f > 1$ values, we check if our analysis is sensitive to the difference between a Standard Model shower and a pure QCD shower. Therefore,

we compare the hypotheses $f = 0$ and $f = 1$. As our null hypothesis in this case expects more events, the distributions of the test statistic is reflected about the y-axis, compared to the standard case when the null hypothesis expects less events. Therefore, the integration that defines the confidence level, is done in the opposite direction (see Appendix A). In Fig. 3.6 we show how different fractional systematic uncertainties σ_{syst} on the bins in the mass distributions limit the exclusion of $f = 0$ from $f = 1$. The confidence level of the exclusion is plotted there as a function of the integrated luminosity. Each row shows the QCD-only hypothesis rejection using one of the methods A-C and each column corresponds to a different dijet p_{T_J} limit. Even though the softer bin $p_{T_J} > 500$ GeV retains the most amount of signal, the S/B ratio is better with a more stringent cut on the fat jet transverse momentum. Therefore, the analysis can exclude the $f = 0$ hypothesis better when performed in a more boosted regime. All three methods allow for a 95% CL exclusion of the QCD-only shower given $\sigma_{\text{syst}} \leq 3.5\%$ in the most boosted bin, but the mass drop and filtering observable m_{BDRS} can exclude it at a much larger confidence level and with a more forgiving uncertainty of $\sigma_{\text{syst}} = 5\%$.

At this point, we stress that this particular modelling of the uncertainty, as a single nuisance parameter with a normal distribution, does not approximate a specific systematic effect. In fact, there are numerous such effects that lead to both normalisation and shape uncertainties. The sources vary from the estimation of the beam luminosity to the object reconstruction efficiency in the various parts of the detector and at different energies. In addition, there are severe theoretical uncertainties associated with the parton distribution functions of the protons and the fixed order calculations of electroweak emissions from quarks. Many of the examples span beyond 5%. However, the purpose of this section is to show how well these systematic effects need to be controlled in order that our methods lead to meaningful statements. We hope that in the course of the LHC lifetime, the various contributions may be parametrised and fitted from other measurements to the desired accuracy.

For the rest of the study we revert back to the more standard situation where the null hypothesis has a lower number of expected events than the alternative

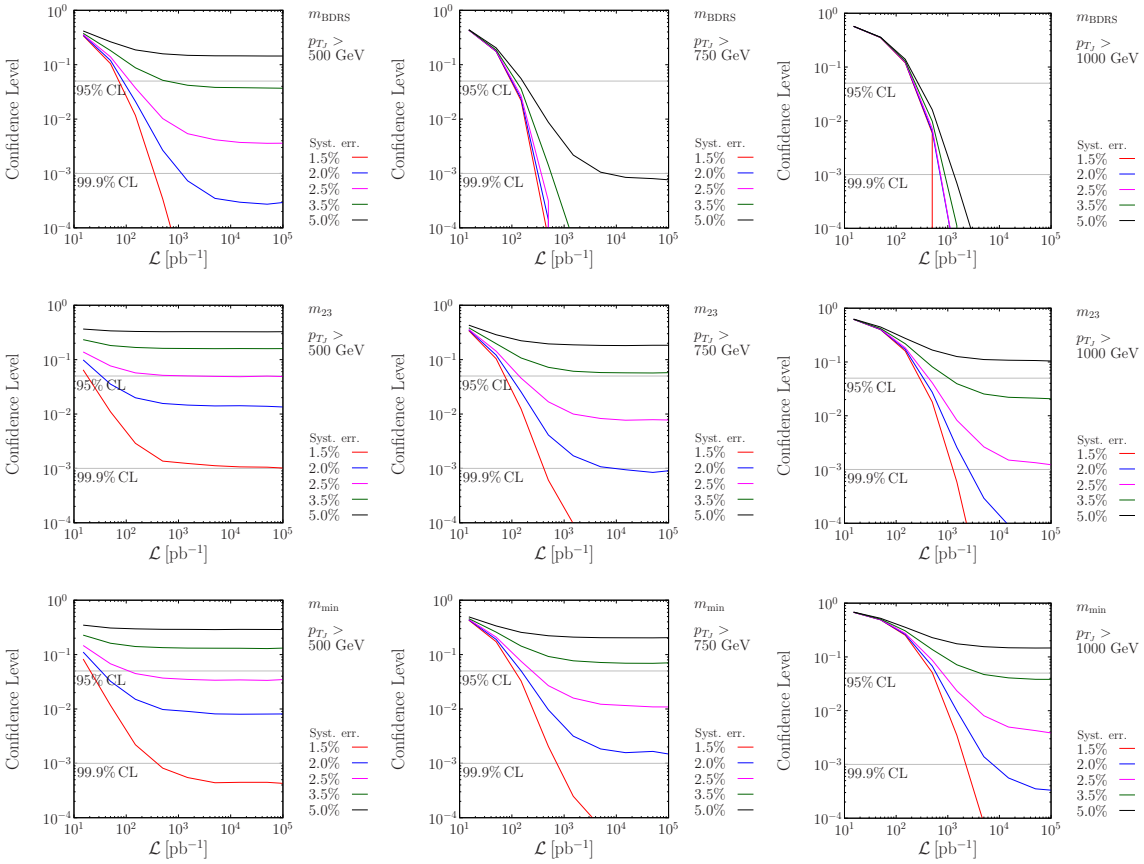


Figure 3.6: $\mathrm{CL_s}$ for the W mass reconstruction through method A using m_{BDRS} (top row), method B using m_{23} (center row), and method C using m_{min} (bottom row) of the hadronic analysis for the three different minimum jet transverse momenta: $p_{T_J} > 500$ (left column), 750 (center column) and 1000 (right column) GeV. The null hypothesis corresponds to $f = 1$ and the alternative to $f = 0$.

hypotheses. In particular we calculate the exclusion confidence level achievable with our analysis of multiplicative factors $f > 1$ given the Standard Model EW splitting probability $f = 1$. All methods retain sufficient number of events after cuts so that the sensitivity of the analysis would be determined by the systematic uncertainty and the S/B ratio at the expected integrated luminosity of the second LHC run. Unfortunately, even a relative uncertainty of 1.5% renders all three hadronic methods A-C insufficient when $f = 1.1$. This is expected as a 10% increase in the W emission rate translates to roughly $\mathcal{O}(1)$% difference between the two hypotheses. We already established that a difference of $\mathcal{O}(10)$%, such as between $f = 0$ and $f = 1$, is detectable by our hadronic methods. Therefore, we expect they would work when

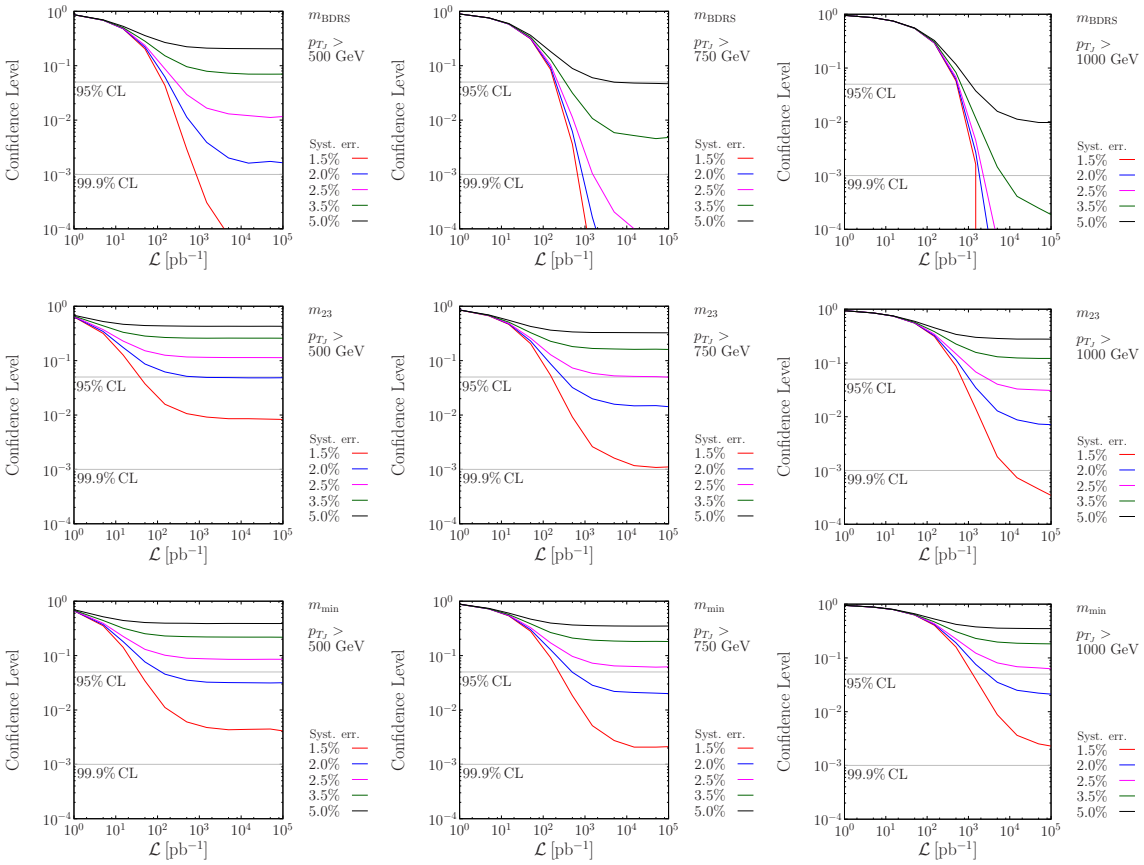


Figure 3.7: CLs obtained from the W mass reconstruction through method A using m_{BDRS} (top row), method B using m_{23} (center row), and method C using m_{min} (bottom row) of the hadronic analysis for the three different minimum jet transverse momenta: $p_{T_J} > 500$ (left column), 750 (center column) and 1000 (right column) GeV. The background corresponds to the Standard Model emission rate ($f = 1$) and signal + background to $f = 2$.

we compare $f = 2$ to the Standard Model. The exclusion confidence level as a function of the luminosity is presented in Fig. 3.7 in the same format as in Fig. 3.6. The exclusion is a little less powerful, but m_{BDRS} can still exclude $f = 2$ at 95% CL with a systematic uncertainty of 5%.

The jet shape variables $\hat{t}$ and τ_{21} were shown to extract additional information from mass-tagged W candidates. Therefore, we can use them to improve the S/B ratio and allow for a more powerful discrimination between the Standard Model and $f = 1.1$. We focus on the strongest method thus far and show in Fig. 3.8 the result of the ellipticity (left) or N-subjettiness ratio (right) applied to the constituents of W-

candidate subjets that pass the mass criterion in method A. Due to the additional mass cut, the statistical uncertainty in the jet shape distributions in the highest p_{T_J} bin remains large even after $\int \mathcal{L}dt = 100\,\text{fb}^{-1}$. Therefore, the plots in Fig. 3.8 are extracted from the bin $p_{T_J} > 750$ GeV, which is boosted enough to allow for an efficient mass reconstruction but also frequent enough to keep the statistical uncertainty under control. The strong QCD rejection at small jet shape values, as discussed at the end of the last section, contributes to a 95% CL exclusion of $f = 1.1$ given a modest systematic uncertainty of $\sigma_{\text{syst}} = 2.5\%$. However, even with this addition the hadronic analysis is not capable of such an exclusion if the uncertainty is 5%.

To do this, we need both a good control over the systematic error and sufficient number of events. The leptonic analysis in Sec. 3.1.2 has a clear advantage when it comes to the systematic error, as the QCD background is irrelevant there. Therefore a $\mathcal{O}(10)\%$ increase in the W emission rate translates directly to a S/B of $\mathcal{O}(10)\%$. As long as the selection cuts are not severe enough to make the statistical error dominant, it is feasible to achieve a 95% CL exclusion of the $f = 1.1$ hypothesis with $\sigma_{\text{syst}} = 5\%$. We can see in Fig. 3.9 that this is indeed the case. One noticeable trend is that not only the statistical uncertainty becomes dominant in the highest p_{T_J} bin, but that the exclusion in the infinite luminosity limit shrinks to slightly lower levels. This is in accordance with the observation that the isolation criterion is affected as the W becomes more collinear. Unfortunately, this is the limit in which the shower treatment will deviate from a fixed order calculation the most. Still, even in the boosted bins, the leptonic analysis allows for the exclusion of $f = 1.1$ if the systematic uncertainty is controlled to 3.5% - 4%.

3.3 Summary of collinear W tagging

We have built a two-step analysis to identify boosted hadronic W bosons, produced in the vicinity of an even harder quark. The emission rate of the electroweak boson in this configuration is increased by Sudakov logarithms $\alpha_W \log^2 Q^2/m_W^2$. Quarks at such a high virtuality are able to produce QCD radiation with sufficiently large

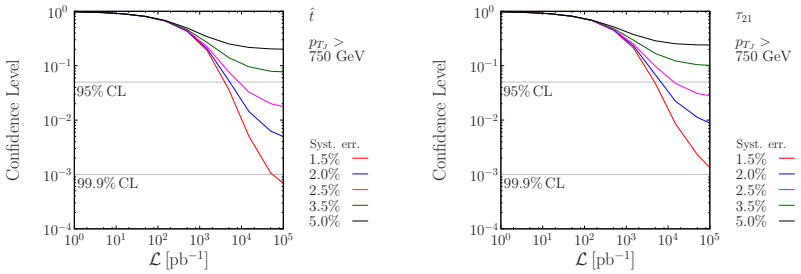


Figure 3.8: CLs obtained from the ellipticity $\hat{t}$ (left) and τ_{21} (right) distributions calculated from the constituents of the W candidates that pass the BDRS cut on the second boosted subjet. $p_{T_J} > 750$ GeV. The background is the SM emission rate ($f = 1$), signal + background sample is $f = 1.1$.

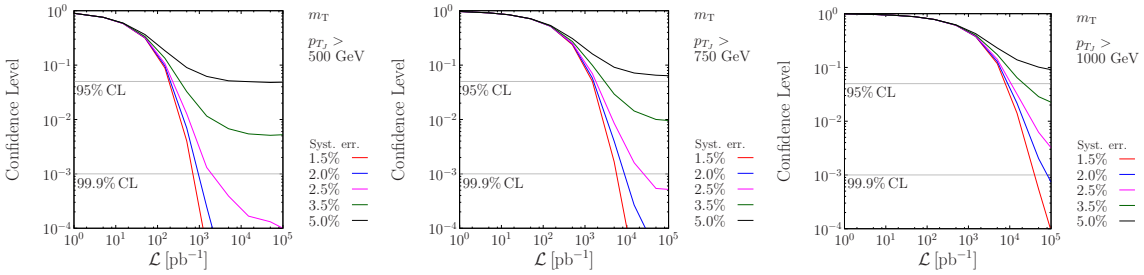


Figure 3.9: CLs obtained from the W transverse mass m_{T} reconstruction in the leptonic analysis. The background sample is the SM emission rate ($f = 1$). The signal plus background sample is $f = 1.1$.

invariant mass to mimic the heavy electroweak bosons. Therefore, we resort to jet substructure techniques to sharpen the mass peak and additional jet shapes that tap into different information about the radiation, such as the colour flow. We see that for a sufficiently low systematic uncertainty of around 5%, mass reconstruction from jet substructure is sufficient to exclude deviations on the order of the expected SM W emission rate, $|\sigma_{f\cdot\mathrm{SM}} - \sigma_{\mathrm{SM}}|/\sigma_{\mathrm{SM}} \approx 1$. To reach sensitivity to 10% deviations, i.e. $f = 1.1$, we include an ellipticity or N-subjettiness ratio cut after the mass reconstruction. Moreover, the analysis can then exclude the $f = 1.1$ hypothesis with an even more strongly controlled systematic uncertainty of 2.5%. After the jet shape cut, the statistical uncertainty crawls back in even after 100 fb^{-1}. Therefore, in the high luminosity run we might expect a better limit. Obviously, the electroweak coupling's strength is well known already, but these types of measurements will

allow for a validation of the effects from the large logarithms associated with the electroweak bosons. Furthermore, it is quite possible that the LHC might not be able to tap fully into the enhanced region, where $m_W^2/Q^2 \to 0$, with sufficient statistics. Therefore, such studies might have to be adapted for a possible future 100 TeV collider.

The QCD background reduction by a single lepton requirement improves the sensitivity greatly. The high statistics bin $p_T > 500$ GeV turns out to be the best for discriminating the $f = 1.1$ hypothesis from the SM $f = 1$ if we search for a leptonic W emission. Even with 5% systematic uncertainty the leptonic analysis is capable of excluding the 10% deviation. Depending on the control of the transverse mass distribution, which is mainly limited by the missing energy $\not{E}$, a leptonic W search may be able to probe even lower deviations.

Chapter 4

Semileptonic $t\bar{t}H(b\bar{b})$

The discovery of the 125 GeV resonance in 2012 [13, 14] shifted the focus of the ATLAS and CMS experiments to identifying the properties of the particle. It is already known to be a spin-0 boson [119] and that its couplings to the Standard Model particles are in agreement with a Standard Model Higgs boson [120]. However, even though the couplings to the heavy gauge bosons have been determined with high precision, the Yukawa couplings of the new particle have still large uncertainties.

In particular, there are important benefits in limiting the uncertainty of the top and bottom quark couplings to the Higgs. The main decay channel of the Standard Model Higgs boson with mass 125 GeV is $H \to b\bar{b}$. Therefore the total decay width Γ_{tot} is dominated by the bottom Yukawa coupling. The cross section of any individual decay channel is proportional to its branching ratio, which involves the total decay width $\text{BR}_i = \Gamma_i/\Gamma_{\text{tot}}$, and implicitly depends on the bottom-Higgs coupling. If it is not constrained, the uncertainty will translate to all coupling measurements [121]. One production mode is $q\bar{q} \to VH$, where V is either a W or Z boson and the Higgs decays to bottom quarks. Because of the two signature mass scales in the final state, and a requirement on the vector boson to decay leptonically, a good signal-to-background ratio can be extracted from this mode [116, 122]. Moreover, the Higgs couples directly both at its production and decay vertex, therefore the extraction of the bottom Yukawa coupling is model independent. Unfortunately, this has not been enough to constrain it as much as the gauge boson couplings during the first Run of the LHC. In the second and third runs, the increased energy and

luminosity open other search strategies for extracting the b-Higgs coupling. One of them is to use the smallest production mode $t\bar{t}H$ [123].

This channel also contains information about the top Yukawa coupling. In the more dominant production channel $gg \to H$, where the Higgs interacts with the gluons through a top loop and therefore depends on y_t, an assumption must be made about what other particles, if any, could contribute to the effective vertex. This is not the case in $t\bar{t}H$, where the vertex is at tree level and the top-Higgs interaction is directly probed. The magnitude of the top interaction is one of the key ingredients in determining the electroweak potential at larger field values, which will tell us how stable the current vacuum state is [124,125]. The other important ingredient is the Higgs self coupling, but the LHC may not be able to provide good estimates of that [126]. Pinning down the top Yukawa is also crucial for the exclusion of various BSM models. For all these reasons, a measurement that can accurately extract the $t\bar{t}H(b\bar{b})$ cross section and contribute to the global fit of the Higgs properties, is worth pursuing.

Both ATLAS and CMS have published analyses specific to the semi-leptonic $t\bar{t}H(b\bar{b})$ channel [127, 128] as well as more general $t\bar{t}H$ searches [129–131] using the data from the first run of the LHC. So far, neither of the collaborations has optimised their reconstruction to boosted phase space regions, but rather both include multivariate (MVA) reconstruction techniques, e.g. boosted decision trees and neural nets, in conjunction with the Matrix Element Method [75].

4.1 Standard Boosted $t\bar{t}H$ Analysis

We update a search performed in 2009 [123] that attempts to reconstruct semi-leptonic $t\bar{t}H(b\bar{b})$ events and distinguish them from a QCD background of the type $t\bar{t}$ + jets and W + jets by exploiting the boosted corner of the final state phase space. The update consists of using more accurate signal and, crucially, background simulations and applying improved reconstruction techniques. The current ATLAS and CMS analyses already vividly show that the S/B ratio is going to be small even in the signal-rich bins. Therefore, a fluctuation in the background model will have a

huge effect on the sensitivity of the analysis. Unfortunately, the discrepancy between the LO simulations in [123] and state of the art NLO shows that the correction to the $t\bar{t}$ + jets channel is of the order of 50%. For a full description of the Monte Carlo simulations used in the analysis see Sec. II of [3]. The new reconstruction method includes a more recent, and now widely accepted, top tagging technique, HEPTop-Tagger [132], which is described in Appendix B. When we present the results, the effects of the change are discussed. In addition, we impose an isolation requirement on the leptons and attempt a more realistic b-jet tagging.

In our analysis we match B-mesons from the MC hadronisation stage to jets and subjets formed with final state objects. If a B-meson falls within the jet radius, then the jet is MC-tagged as a b-jet. When all jets and subjets in a configuration are MC-tagged as b-jets or light jets, a b-tag weight is given to the configuration as a whole. This happens by calculating the probability to find a fixed number of b-jets and light jets from the experimentally quoted efficiencies (70% and 1% respectively for MC-tagged b-jet and light jet). The approach we adopt does not exactly simulate the experimental method, but is conservative in so far as we do not correct for the energy of invisible decay products of B-mesons, which will result in a smeared out m_{bb} distribution and thereby reduce the statistical sensitivity of our reconstructions. Previously the hadrons in the final state were not decayed, so a jet could actually contain a B-meson in its constituents, which simplifies the procedure, but introduces an unrealistic energy resolution. The origin of this is that a B-meson has a decay channel of with missing energy. Since the resonance we are looking for contains two kinematically significant B-mesons from the $b\bar{b}$ pair, reconstructing the resonance after hadronic decays will smear and displace the peak.

The analysis is performed with three types of objects: *hadrons, leptons* and *B-mesons*. The leptons are associated with $\ell \in \{e^{\pm}, \mu^{\pm}\}$ and include isolation and kinematic criteria. To consider ℓ as a *lepton* from the hard interaction and heavy objects' decays we require that it is central $|\eta_\ell| < 2.5$, sufficiently energetic $p_{T\ell} > 25$ GeV [128] and isolated from the hadronic radiation $\sum_{i \in \Delta R_{i\ell} < 0.2} H_{T\mathrm{i}} < 0.1\, p_{T\ell}$. A *hadron* is any other visible final state particle with the more relaxed kinematic constraints $|\eta| < 4.5$ and $p_T > 0.5$ GeV. Finally, the B-mesons are not directly

involved in any reconstructed object. However, they provide the means to assign b-tags to jets and subjets as discussed in the previous paragraph. In order to qualify for b-tagging, a B-meson must satisfy the following kinematic constraints: $p_T > 10$ GeV and $|\eta| < 2.5$. Consequently, a jet or subjet is only viable for b-tagging if it also satisfies the same pseudorapidity constraint.

The analysis begins with simple selection cuts that eliminate the overwhelming QCD background, which is not included further. The first requirement is of a single isolated lepton. Because such a lepton originates from the matrix element (or EW scale resonance decays), this condition eliminates the pure multi-jet QCD background. It also separates this analysis from the fully leptonic $t\bar{t} + \mathrm{X}$ processes, which are not the subject of this study. The hadrons in events that pass the first requirement are clustered into CA fat jets with $R = 1.5$ and $p_{Tj} > 200$ GeV. The second selection cut is of at least two such fat jets, from which the hadronic top t_{had} and Higgs boson will be extracted. The transverse momentum limit is not very large. In fact it is only slightly above the top mass. Therefore, the particles we are looking for will be only slightly boosted. Usually, a stronger boost benefits the S/B ratio; however, as we are dealing with the least frequent Higgs production channel, we have to keep in mind the signal efficiency and not only the purity of the sample. Therefore, we cannot afford large boosts in this search.

After the two event selection cuts in the previous paragraph, the stage is set for the actual jet-substructure analysis of the hadronic top and Higgs candidates. The reconstruction of the event proceeds in seven steps.

1. The HEPTopTagger is applied to each fat jet and as a result each fat jet is either tagged as t_{had} or non-t_{had}. Usually, a semi-leptonic $t\bar{t}$ event is rarely going to receive double t_{had} tag, but the second hadronic resonance in the event actually substantially increases the odds (see Sec. 4.1.1). Therefore, we are forced to drop multi-t_{had} events or select a "best" top.

2. In the interest of retaining as much signal as possible, we choose the second path, instead of vetoing such events, by selecting the top candidate that minimises $\Delta m_{\mathrm{tot}} \equiv |m_{\mathrm{t,reco}} - m_{\mathrm{t}}| + \min_{\mathrm{ij}} |m_{\mathrm{ij}} - m_{\mathrm{W}}|$. Here $m_{t,\mathrm{reco}}$ is the mass of the

reconstructed top and m_{ij} is the invariant mass of the pair of subjets closest to the W mass. This top candidate jet is ignored for the rest of the event manipulation.

3. For b-tagging purposes a rapidity cut $|\eta| < 2.5$ is applied to all remaining fat jets, including top-tagged jets that have not been selected as the event top candidate in the previous step.

4. Each of the remaining fat jets (usually only one) goes through the mass drop filtering procedure proposed in [123]. If the mass drop leaves only one subjet we move to the next fat jet (or reject the event if all satisfy the condition). Otherwise the pairs of 4-momenta that survive the mass drop represent possible $H(b\bar{b})$ structures. At this point it is possible to fall into a combinatorial problem from all the possible pairs, which is the opposite of what a boosted analysis relies on. To avoid it, the pairs are ordered according to the distance,

$$d_{\mathrm{ij}} = p_{\mathrm{Ti}} p_{\mathrm{Tj}} \Delta R^4_{\mathrm{ij}}, \tag{4.1.1}$$

and only the first three such pairs in descending distance d_{ij} are retained. The constituents of each remaining pair are filtered into C/A jets of radius $R_{\mathrm{filt}} = \min(0.3, \Delta R_{\mathrm{ij}})$ and $p_{\mathrm{T}} > 20$ GeV. Only the hardest 3 filtered jets are kept and combined into what we refer to as a *Higgs candidate*.

5. We require exactly two b-tags from the filtered subjets of the Higgs candidate.

6. An additional b-tag can be applied in order to combat $t\bar{t}$ + light jets background. To do so we use all hadrons in the event that are **not** already in the t_{had} or Higgs candidate. If the event structure is correctly reconstructed, there should be only a single b-quark among them. There are two independent sets of hadrons, which we treat separately. One set contains the hadrons not in either the top fat jet or the fat jet with the Higgs candidate. They are clustered into C/A jets with $R = 0.4$ and $p_{\mathrm{T}} > 30$ GeV, which we call *outer jets*. The other set consists of the hadrons in the fat jet that contains the Higgs candidate, but not contributing to the candidate itself. They are reclustered into C/A jets with $R = R_{\mathrm{filt}}$ and $p_{\mathrm{T}} > 20$ GeV - *inner jets*. As the Higgs

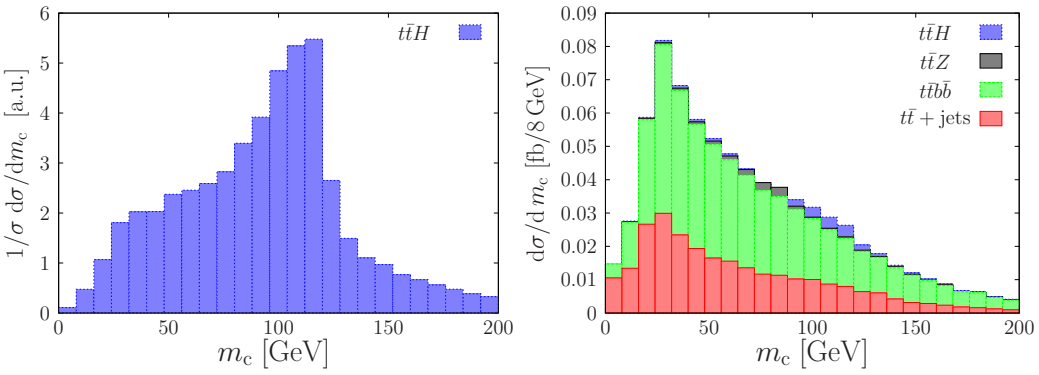


Figure 4.1: Distributions in the Higgs-candidate mass, m_c, for signal (left) and signal plus $t\bar{t}+X$ backgrounds (right) after step 6 (third b-tag) of the standard boosted analysis of Sec. 4.1.

fat jet was already processed by a mass drop/grooming procedure, we choose a smaller jet definition. From the combined set of inner and outer jets, we request a single b-tagged jet to continue.

7. Finally, we identify a Higgs candidate as tagged if its invariant mass m_c lies in the $[100, 130]$ GeV mass window.

We see that the 6-step analysis (before the last mass cut) leads to a mass distribution m_c with a resonance signature, but also some undesirable features (left plot in Fig. 4.1). Going from the large-mass end of the distribution to lower values, there is a significant and sharp peak. Unfortunately, on the low-mass side of the peak there is a much more slowly decreasing tail that merges into a bulge around 50 GeV. This underlying structure under the peak comes from mistagged Higgs candidates. Usually it happens when a Higgs candidate is polluted by b-quarks from tops. The Higgs peak is also shifted by about 15 GeV to the left of where it is supposed to be because even in a correct Higgs identification there are two B-mesons that often decay leptonically and the neutrinos carry away energy and mass from the containing jet. The right plot in Fig. 4.1 shows how the Higgs mass distribution compares with that from the dominant background processes of type $t\bar{t}+\mathrm{X}$. Even though the bulk of the background is concentrated in the low-mass region, as expected from low-mass gluon splitting, there is still significant irreducible background left in the region of the reconstructed Higgs resonance. The result will be analysed more in

depth in Sec. 4.4, but it should be noted that the outcome of the reconstruction is worse than what was found in [123]. The S/B ratio is lower due to a combination of the new b-tagging smearing of the signal over a wider range of masses and the NLO normalisation, which results in a direct increase in the background. This opens up an old problem of this Higgs production mode, which [123] seemed to eradicate. The theoretical and experimental systematic uncertainties in our control over the background distributions may easily be comparable to the signal. We need a better background reduction strategy and a better handle on the uncertainty if this channel is to be of any use.

We do not provide any suggestions for the latter, but in the next section we attempt to improve the former. First, we look exactly how well our event reconstruction works out. In particular we evaluate how often our reconstructed candidates match the matrix element particles they are supposed to represent. We look into what matrix element particles form the different fat jets and what particle-jet configurations (which we call *event topologies*) contribute most to the different parts of the mass distribution. Some of these event topologies are shown in Fig. 4.2. The logic of the reconstruction analysis so far stems from associating the modestly boosted $t\bar{t}H(b\bar{b})$ event with the topology in Fig. 4.2a. The Higgs decay products are well spatially separated from the hadronic top products and leptonic top b-quark is in neither of their vicinities. However, the rest of the figure contains only three of the numerous possible combinations. The sheer number of particles involved at tree level, means that it is very easy to have unboosted tops and Higgs, but still be able to reconstruct two 200 GeV fat jet out of random combinations of their decay products. A more boosted requirement will sift out the unwanted topologies, but we have to balance that with the expected available signal at the LHC. In the next subsections we focus on the signal event topologies that contribute most to the smeared m_{c} distribution and spoil the quality of the Higgs peak.

4.1.1 Quality of hadronic top reconstruction

We measure the proximity of a hadronic top candidate fat jet to the ideal topology with the following 8 binary conditions (true/false):

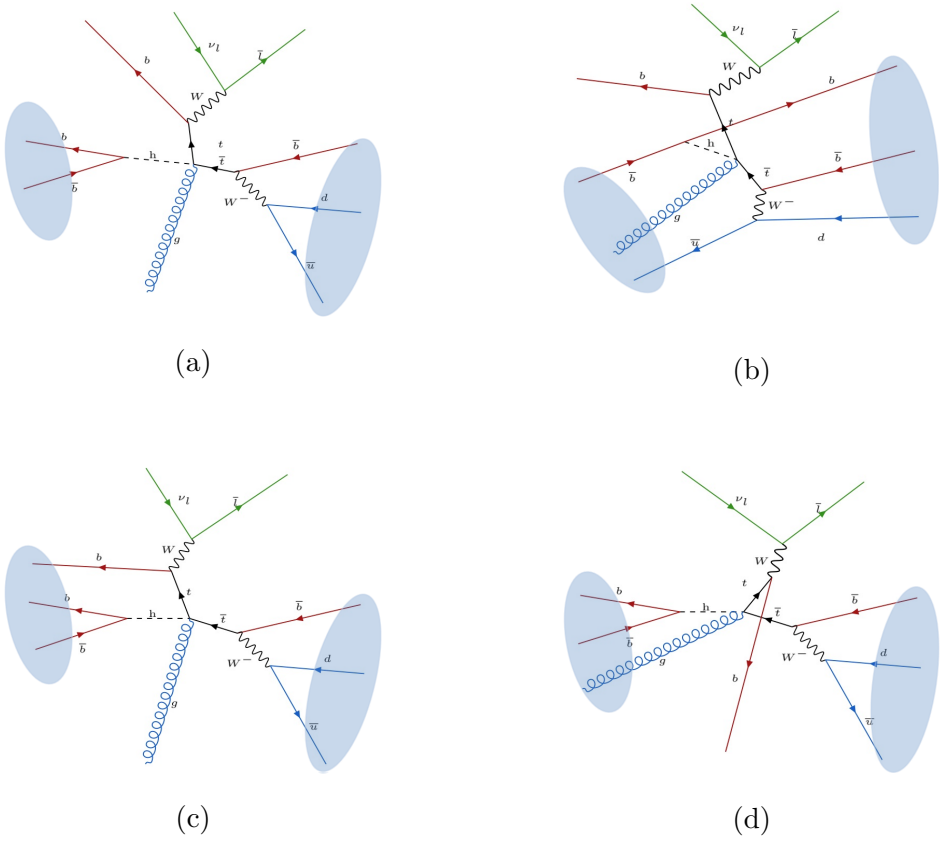


Figure 4.2: Schematic representation of typical $t\bar{t}H$ event topologies. The ellipses indicate how partons are clustered to form two fat jets. Topology 4.2a is the cleanest one: the Higgs products and the hadronic top products form two separate fat jets without pollution from other hard particles. Topology 4.2b features misassignements of the Higgs and hadronic top products. In topology 4.2c the hadronic top decay products form a fat jet, and the Higgs decay products form another fat jet with the leptonic top b-quark falling within it. In topology 4.2d the b-quark from the leptonic top decay does not pollute the Higgs fat jet, but there is a gluon radiation strong enough to form a substructure within the Higgs fat jet.

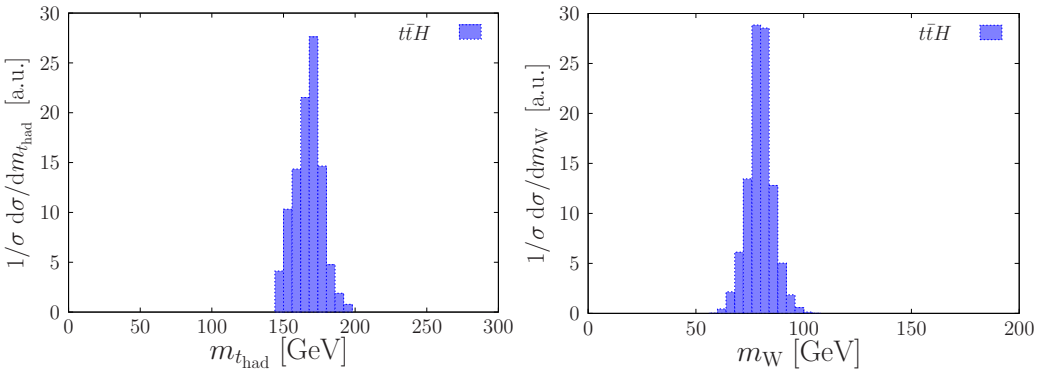


Figure 4.3: Distributions of the $m_{\mathrm{t_{had}}}$ (left) and m_{W} (right) invariant masses for the cleanest topology A_1 of Table 4.1, after step 2 of the boosted analysis of Sec. 4.1.

1. t_{had}: the hadronic top quark is boosted ($p_{\mathrm{T,t_{had}}} > 150\,\mathrm{GeV}$)
2. t_{had}: the hadronic top quark overlaps with the jet ($\Delta R_{\mathrm{jet,t_{had}}} < R_{\mathrm{fat}}$)
3. $t_{\mathrm{lep}} \to b\ell\nu$: the b-quark from t_{lep} belongs to the jet
4. $H \to b\bar{b}$: the harder b from the Higgs belongs to the jet
5. $H \to b\bar{b}$: the softer b from the Higgs belongs to the jet
6. $t_{\mathrm{had}} \to bjj$: the b-quark from t_{had} belongs to the jet
7. $t_{\mathrm{had}} \to bjj$: the harder light quark from t_{had} belongs to the jet
8. $t_{\mathrm{had}} \to bjj$: the softer light quark from t_{had} belongs to the jet

Each fat jet, characterised by these binary variables, falls into one of a total of 256 possible bins. We refer to these bins as *jet topologies*. The evaluation is done at two stages of the analysis. The first is just before a top candidate is selected (step 1) and the other is right after that. At the second stage we evaluate the topology of the unique t_{had} jet. The number of possibilities would make the task of analysing the topologies too difficult. Thankfully, more than 60% of all fat jets that have been identified as the hadronic top at step 1 fall into one of the six topologies in Table 4.1.

The topology in the first row corresponds to the ideal scenario. We have a boosted hadronic top in the direction of the t_{had} and all three of its decay quarks fall within the radius of the fat jet. Moreover, neither of the remaining b-quarks in

label	bin	before top tag	after top tag	tagging efficiency
A_1	11000111	0.12	0.32	0.40
A_2	11001111	0.03	0.08	0.42
A_3	10111000	0.06	0.07	0.18
A_4	11010111	0.02	0.06	0.40
A_5	11100111	0.02	0.04	0.41
A_6	11011111	0.01	0.04	0.39

Table 4.1: The normalised distributions of fat jets before top tagging (column 2) and top-tagged fat jet (column 3) in the dominant bins of the 8-dimensional jet-category histogram. The top-tagging efficiency (column 4) is defined as the probability that a fat jet is top-tagged in step 2 of the boosted selection. The rows are ordered by decreasing fraction after the top-tag. The bin is identified by specifying the conditions that are true (1) and false (0) in the order listed in the text. The left-most digit corresponds to the first condition.

the event contaminate the jet. Unsurprisingly, the mass reconstruction of both the hadronic top and the associated W boson is very clean as testified by the plots in Fig. 4.3. Even though we use the knowledge from the Monte Carlo event record, these two peaks are reconstructed from final state particles and therefore carry all the smearing from parton shower, hadronisation, initial state radiation and multiple parton interactions within the protons. This goes to show the effectiveness of a top tagger[1] in removing spurious radiation while preserving the hard structure in a fat jet. The second column entry of topology A_1 in Table 4.1 confirms the assertion that the picture that guides the steps of the analysis is not full. This perfect configuration occurs only in a quarter of the $t\bar{t}H(b\bar{b})$ events before top-tagging and a third after top-tagging. The only other topology that does not involve any Higgs decay products is when the b-quark from t_{lep} ends up in the top fat jet A_5. We see that the reconstruction efficiency is 40%, just as in the purest case. These are the only two

[1]The distributions in Fig. 4.3 are obtained from the top tagger employed in [123] because it has a designated W candidate.

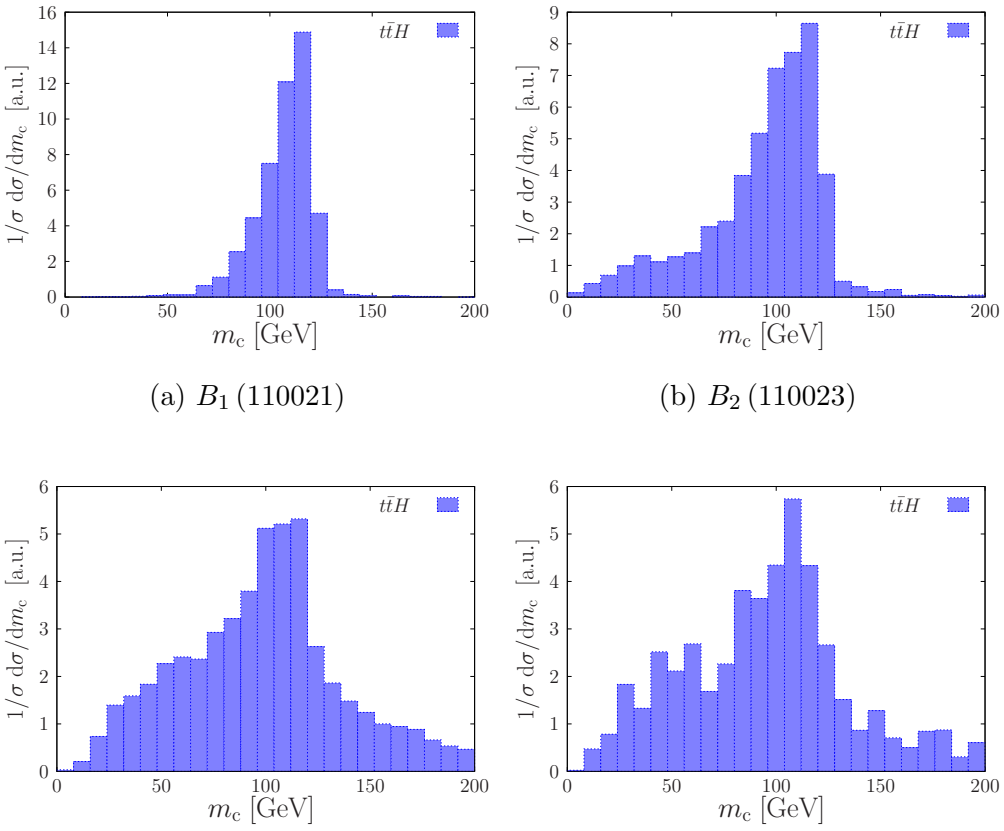
top topologies that would allow the true Higgs to be identified at a later stage. A curious feature is that the same tagging efficiency is accomplished even when some of the Higgs decay products pollute the top, as long as all the three t_{had} quarks are involved in the jet. This means that the top tagger manages to sift through all hard subjets in a fat jet and isolate the ones that form the best top candidate.

The A_3 topology is qualitatively different from the rest. This is when all the wrong b-quarks (two from the Higgs and one from the leptonic top) form a single fat jet with the correct mass structure between them, so that the HEPTopTagger identifies the configuration as a top, even though the true top is in the opposite side of the detector. These configurations are top-tagged with a 20% efficiency, which is rather substantial, especially compared to pure QCD mistag rate. At first thought, such a misidentification seems to make Higgs tagging impossible. But actually it also means that a pure top jet lies in the opposite direction in the same event and will be tagged as well with a 40% efficiency. Therefore, our decision not to veto events with multiple hadronic top tags, but select the best out of them, contributes to more signal retention without affecting the background processes, as there is no third resonance in $t\bar{t}$ + jets that would fake a top in combination with the t_{lep} b-quark. The topologies with a mix of top and Higgs products (A_2, A_4, A_6) on the other hand can never lead to a correct Higgs reconstruction. All in all similar configurations amount to more than 50% of all events after top-tagging (this number includes all 256 topologies and not just the six in Table 4.1). Fortunately such configurations invest too many of the b-quarks into t_{had}, so not enough jets and subjets will get the needed b-tag. Therefore, such false Higgs configurations are naturally vetoed by the analysis.

4.1.2 Quality of Higgs reconstruction

Similarly to the hadronic top jet, we examine how the Higgs candidate fat jet forms from the hard interaction particles by classifying the jet according to several conditions.

1. H: the Higgs boson is boosted ($p_{\text{T,H}} > 150\,\text{GeV}$)



(a) B_1 (110021) (b) B_2 (110023)

(c) B_3 (110123) (d) B_4 (111023)

Figure 4.4: Distributions of the Higgs candidate mass, m_c, for different Higgs-jet topologies after requesting three b-tags, i.e. after step 6 of the boosted analysis. The figures correspond to the topologies shown in Table 4.2.

label	bin	before b-tags	after b-tags	after m_c cut	tag efficiency
B_1	110021	0.05	0.08	0.17	0.77
B_2	110023	0.10	0.16	0.24	0.53
B_3	110123	0.09	0.40	0.38	0.32
B_4	111023	0.01	0.03	0.03	0.31

Table 4.2: The fraction of the signal cross section at different steps of the analysis in four of the 144 bins in the 6-dimensional Higgs-jet category histogram. The tag efficiency of the topology is reported in the last column, and the bins are ordered by decreasing tag efficiency. Each row corresponds to a bin identified by specifying the conditions that are true and false (or a numerical value if applicable) in the order listed in the text. The left-most digit corresponds to the first condition.

2. H: the Higgs boson overlaps with the jet ($\Delta R_{\text{jet,H}} < R_{\text{fat}}$)

3. $t_{\text{had}} \to bjj$: the b quark from t_{had} belongs to the jet

4. $t_{\text{lep}} \to b\ell\nu$: the b quark from t_{lep} belongs to the jet

5. $H \to b\bar{b}$: the number of b-quarks from the Higgs decay the jet contains is 0/1/2

6. $H \to b\bar{b}$: the number of $b\bar{b}$ Higgs candidates in the fat jet is 0/1/3

Categories $(1-4)$ are very similar to the top classification categories and just like them have a binary outcome. The last two categories on the other hand have three possible values. The Higgs candidate selection step 4 is such that there can be a very limited number of Higgs candidates per fat jet depending on the substructures after the mass drop. In the case that no subjets remain there are no candidates. If there are two, they form a single pair. Three subjets can be paired in three ways and for any larger number we select the top three pairs according to the distance d_{ij} defined in the step. That is why the only three outcomes for condition 6 are 0/1/3. All in all our classification of Higgs fat jets consists of 144 independent jet topologies. Here the signal after the Higgs identification stage is even more concentrated in only a few topologies. In Table 4.2 we show the contribution of four topologies at three

steps in the analysis in the beginning of the section: before b-tagging (before step 5); after b-tagging (after step 6); after the final cut on m_{c} (step 7). In the final step these four topologies contribute to 80% of the total signal contained in the m_{c} cut. In Fig. 4.4 we show the mass distributions of the four topologies after all b-tagging (step 6) but before the final mass cut. Again the characteristic B_1 topology, with both b-quarks from the Higgs decay falling within the fat jet and no other EW resonance decay products contaminating the jet, provides the purest Higgs peak. There is no way around the missing energy from the B-meson neutrino slipping away from detection. The resonance is both skewed towards lower mass values and its peak is shifted down. Adding an additional strong QCD subjet (B_2 topology) that would not be removed in the mass drop procedure slightly changes the mass distribution because each fat jet now contains three Higgs candidates and only one has the true mass scale. The other candidates will contribute to a background-like mass distribution with the bulk of the cross section at smaller masses. In this case only the true candidate has two b-tags, therefore the contribution from the false candidates is diminished.

The largest contribution to the final cross section comes from the topology B_3, where in addition to the b-quarks coming from the Higgs decay, the fat jet also captures the leptonic top b-quark. This acts very much like a third QCD subjet when it comes to the shape of the mass distributions from different candidates, but the difference is that the wrong candidates are no longer suppressed by a lack of b-quarks. The same argument is true in the B_4 topology, where the t_{had} b-quark ends up among the Higgs remnants. However, the contribution of B_4 is negligible because the b-quark is vital in the top identification at an earlier step of the analysis. It is rare that the two mass scales of a hadronic top will be correctly mimicked, leading to discarding such topologies by the HEPTopTagger before they reach the Higgs tagging stage. Going back to the significant B_3 topology, the background-like contribution to the mass distribution from Higgs mistags happens only because we cannot distinguish which of the b-tagged subjets originates from the leptonic top. If we are able to simultaneously reconstruct t_{lep} and the Higgs, the ambiguity will disappear and the Higgs peak will sharpen.

4.2 Improvements and new avenues

As a continuation from the previous section 4.1 and following the discussion in 4.1.2 of the major Higgs misidentification topologies, we present an augmented Higgs search strategy. Moreover, we expand the phase space region to include single boosted fat jet events as well. At the end of the section, for comparison with the combined sensitivity of all our independent search strategies, we do a simple MVA on a phase space that does not necessarily contain boosted objects.

4.2.1 Boosted final state configurations

The analysis steps in the previous section (4.1) target a very specific event topology - the combination of A_1 and B_1 in tables 4.1 and 4.2 respectively. In those cases the numerous resonances in the event are well reconstructed. The difficulty lies in correctly identifying the different particles when the topology does not match the ideal scenario. After the results in Sec. 4.1.2, we are in a position to separate the pure topology B_1, which already gives the sharp peak needed for a successful reconstruction of the signal event, from the dominant topology B_3, which requires additional work. Most of the classification parameters that define the topology of a jet rely on Monte Carlo information and cannot be used directly in an experimental analysis. One condition is an exception. We can safely separate the Higgs fat jet into two categories according to the number of Higgs candidates within the jet. The topology B_1 happens when the mass drop leaves only two subjets and a single Higgs candidate, while the more troublesome B_3 topology happens when the mass drop procedure allows more than two subjets to remain. Therefore, we can treat the two-subjet and multi-subjet cases independently. B_1 is not the only topology that can contribute to a two-prong fat jet. For example a fat jet that contains the b-quark from t_{lep} and one of the two Higgs b-quarks can also end up in this category. However, the final mass distribution of the Higgs candidate in two-prong fat jets is heavily dominated by B_1. It is the only topology of this type in the top four contributors to the final signal. Therefore, we can assume a successfully tagged Higgs candidate from a two-prong fat jet is always the true Higgs.

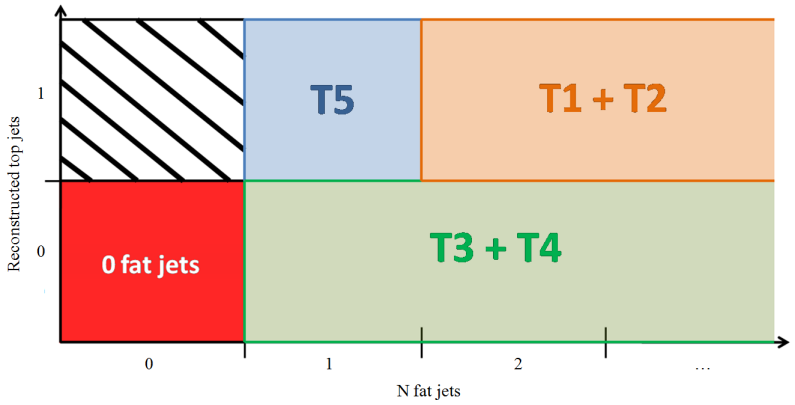


Figure 4.5: The single-isolated-lepton event phase space with the explored regions labelled as in the text.

In the following we augment the analysis in Sec. 4.1 when we deal with multi-prong fat jets in order to alleviate the Higgs mistag rate in $t\bar{t}H$ events with more complicated topologies. Moreover, to the benefit of the statistical significance of the search, we analyse extra, statistically independent, selection channels where we treat either the Higgs or the hadronic top as boosted but not the other. A simple diagram of how the different search strategies fit into the $t\bar{t}H(b\bar{b})$ phase space is shown in Fig. 4.5:

T1: ≥ 2 fat jets, 1 tagged boosted top, 1 Higgs candidate

T2: ≥ 2 fat jets, 1 tagged boosted top, 3 Higgs candidates

T3: ≥ 1 fat jets, no tagged boosted tops, 1 Higgs candidate

T4: ≥ 1 fat jets, no tagged boosted tops, 3 Higgs candidates

T5: exactly 1 fat jet, 1 tagged boosted top, unboosted Higgs candidate

Note that it is possible to separate all five configurations in bins of their own, but we need an additional direction for the number of Higgs candidates within a fat jet. For clarity, this direction is integrated out in the diagram of Fig. 4.5. Nevertheless, in our analysis all five are statistically independent. The first two configurations, **T1** and **T2**, represent the entire phase space region in the original analysis of Sec. 4.1. From now on we treat them differently. The categories **T3** and

T4 look for configurations where the Higgs boson is boosted, but the hadronic top is not. Finally, **T5** is concerned with an unboosted Higgs boson after a boosted t_{had} has been identified. We leave the strictly unboosted bin out of the analysis, but it will be incorporated into the MVA search later.

Topologies T1 and T2: Boosted $\mathbf{t_{had}}$ and boosted H

We focus first on separating the original analysis of Sec. 4.1 into two individual searches for the cases of one Higgs candidate (**T1**) and three Higgs candidates (**T2**) within a fat jet.

The mass reconstruction in the **T1** channel is already quite successful. Therefore, in order to improve the S/B ratio between $t\bar{t}H$ and $t\bar{t}+X$ backgrounds we need to look for other differences. For example, we expect a different colour structure between the decay products of a colour singlet, as is the case with the Higgs boson, and the dominant background $t\bar{t}b\bar{b}$ where the $b\bar{b}$ pair usually originates from a gluon. For those cases the colour dipoles that the b quarks form are very different. The Higgs b-quark pair forms a single dipole, which disfavours any further QCD radiation at a large angular distance from the cone defined by the two quarks. On the other hand, each quark from the background $b\bar{b}$ pair forms such a dipole with a different particle. Such a physics signature was already used in Chapter 3 via jet shape observables. Here we apply the ellipticity $\hat{t}$ to the Higgs candidate constituents. As already pointed out, the bulk of the signal distribution is clustered at low values. Therefore, we can compare how an ellipticity cut $\hat{t} < 0.2$ changes the mass distribution m_{c} (see Fig.4.6). In the results section Sec. 4.4 we show that this cut improves substantially the S/B ratio at small cost to the overall signal retention. However, this channel is intrinsically rare and **T1** does not provide sufficient statistical sensitivity to be used individually in Run 2.

The channel **T2**, which is dominated by the jet topology B_3, occurs four times more frequently than **T1**. Therefore, getting a better S/B ratio in this channel is key to improving the sensitivity to $t\bar{t}H(b\bar{b})$ events over other $t\bar{t}+X$ backgrounds. The important conclusion from Sec.4.1.2 is that the signal is smeared due to the indistinguishable Higgs and t_{lep} b-quarks. Finding a method to make them distin-

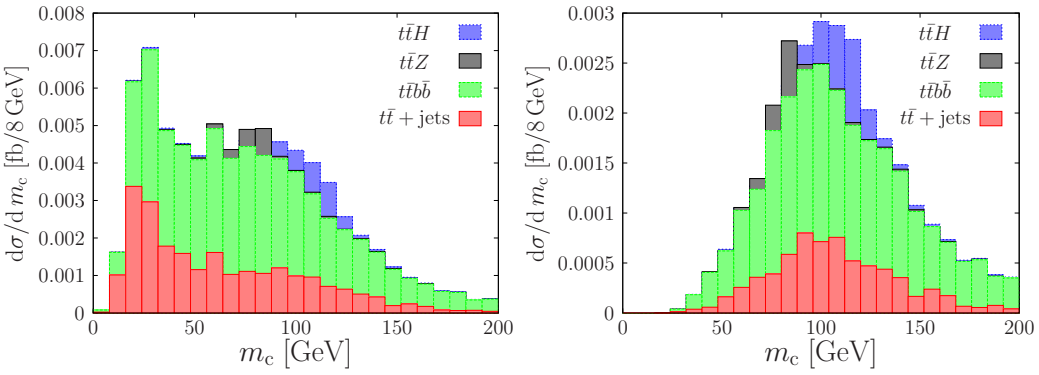


Figure 4.6: $m_{\rm c}$ distribution from the selection channel with a single Higgs candidate in the fat jet and a tagged boosted hadronic top (**T1**). The left(right) figure is without(with) a $\hat{t}$ cut on the Higgs candidate constituents.

quishable can force the signal into the Higgs peak region. Therefore, we attempt to tag both the Higgs and $t_{\rm lep}$ after step 4 of the original analysis of Sec. 4.1 in the case when the Higgs fat jet contains three Higgs candidates. Once a leptonic top is identified, there is only one pair of b-tagged subjets that can form a Higgs and the combinatorial bulky structure in Fig. 4.4c will be removed. The reconstruction of the Higgs and leptonic top is done simultaneously by minimising a χ^2 variable, which is computed for each combination of final state objects that form a possible reconstructed Higgs-$t_{\rm lep}$ pair. We already established that there are three Higgs candidates in the fat jet. Each of these candidates is associated with multiple combinations that can form the top. The physical objects involved in the reconstruction are:

1. two subjets reconstructed from the hadrons of the filtered Higgs candidate using the exclusive-k_T algorithm.

2. the inner and outer jets with respect to the current Higgs candidate (see definition in Sec.4.1);

3. the isolated lepton;

4. the missing transverse momentum of the event $\not{E}_{\mathbf{T}}$.

The event has only a single missing particle, the neutrino from the leptonic W decay. Therefore, through the conservation of the transverse momentum, in theory

there is only one degree of freedom corresponding to the neutrino momentum in the beam direction. It can be constrained from the energy equation of an on-shell W boson, the lepton momentum and the missing transverse momentum. Since the relativistic energy equation is quadratic, there are two potentially different roots for the neutrino p_z component. This ambiguity is not resolved here, but each root is associated with a separate configuration awaiting a χ^2 value. In addition to the Higgs candidate we need a leptonic top, which consists of a b-quark, a charged lepton and a neutrino. Therefore a Higgs-t_{lep} configuration is any unique choice of one of n inner and outer jets, one of the two neutrino candidates, the isolated lepton, and the two exclusive Higgs candidate subjets. Thus, any fat jet with 3 Higgs candidates has a total of $2\sum_{i=1}^{3} n_i$ configurations. And each of these configurations gets a χ^2 score defined by

$$\begin{aligned}
\chi^2 &= \chi^2_{\text{top}} + \chi^2_{\text{Higgs}}, \\
\chi^2_{\text{top}} &= \frac{(m_{\text{t}_{\text{lep}},\text{reco}} - m_{\text{t}_{\text{had}},\text{max}})^2}{\sigma^2_{\text{t}_{\text{had}}}}, \\
\chi^2_{\text{Higgs}} &= \frac{(m_{\text{H,reco}} - m_{\text{H,max}})^2}{\sigma^2_{\text{H+}}}\Theta(m_{\text{H,reco}} - m_{\text{H,max}}) \qquad (4.2.2) \\
&+ \frac{(m_{\text{H,reco}} - m_{\text{H,max}})^2}{\sigma^2_{\text{H-}}}\Theta(m_{\text{H,max}} - m_{\text{H,reco}}), \qquad (4.2.3)
\end{aligned}$$

where Θ is the Heaviside step function. The errors $\sigma_{\text{H}\pm}$ are the standard deviations of Gaussian fits to the data to the right (+) and left (-) of the peak in **T1** (Fig. 4.4a). The reason for two Gaussian fits is that even the purest Higgs peak is so significantly skewed, that a single averaged value cannot be an accurate description of the spread of the peak in one or the other direction, and quite possibly both. Naturally $m_{H,\text{max}}$ is the position of the peak. To extract the parameters associated with the top, we use the t_{had} distribution obtained from the purest topology A_1 (left plot on Fig. 4.3). The resonance is symmetric enough that a single Gaussian fit can suffice to extract $\sigma_{\text{t}_{\text{had}}}$ and $m_{\text{t}_{\text{had}},\text{max}}$. Once all configurations get an associated χ^2 score according to Eq. (4.2.2), they are sorted by ascending χ^2. Only the first quarter of unique Ht_{lep} configurations are kept. Of those we require two b-tagged subjets in the Higgs candidate and another b-tag for a single inner or outer jet (remember that each unique configuration has only one such jet associated with it). We record the Higgs

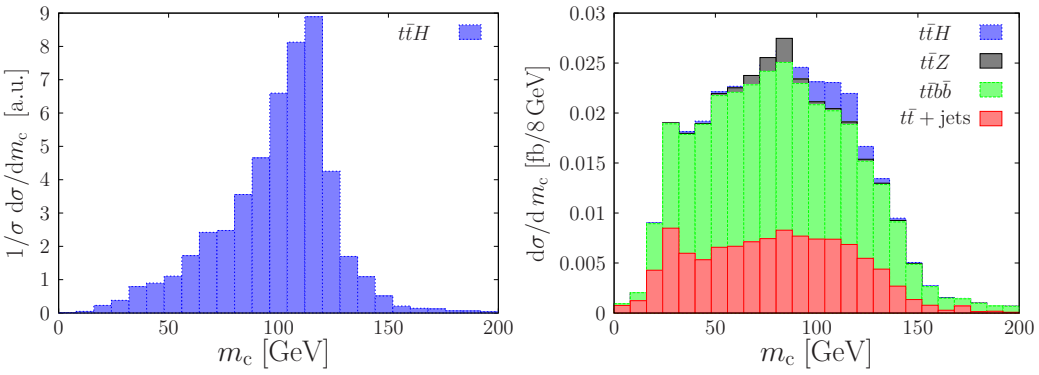


Figure 4.7: m_c distribution obtained from the 25% of configurations with lowest χ^2 score in the 3-Higgs-candidate selection channel (**T2**). The left figure is the signal $t\bar{t}H$ and the figure to the right contains signal and background.

candidate mass m_c from all configurations that successfully pass both the χ^2 and b-tag cuts.

At first glance it seems like the new addition to the analysis in Sec. 4.1 takes a three-fold combinatorial problem and makes it worse. However, the χ^2 cut complements the b-tagging. Previously we had three Higgs candidates contributing with equal weight to the final mass distribution, but only one of them was the true Higgs. Now, even though there are multiple configurations, there are still only three that involve all of the three b quarks. All the rest, even if they end up in final distribution, will be severely diminished by the $3b$-tag requirement. Therefore, in principle, as long as the χ^2 score keeps the true Ht_lep configuration among the top 25% more often than the other two significant (but false) configurations, the reconstruction of the Higgs mass should improve. To see the effect of the procedure, compare the distribution m_c from the new method in left plot of Fig. 4.7 to the mass distribution from topology B_3 extracted from the original analysis (Fig. 4.4c).

Just as in the simpler case of channel **T1**, we can use other physical arguments beyond mass reconstruction, in order to attempt to remove the false Higgs configurations before the b-tagging step. We attempted to use the colour structure of the colour singlet Higgs with $\hat{t}$ and also applied a cut on the helicity angle of the leptonic top b-quark [133]. Unfortunately neither of them contributed in any meaningful way in increasing S/B in this channel.

It should be noted that at this point we have reconstructed all of the matrix element objects in a $t\bar{t}H$ event. Therefore we can look for angular dependencies between these fundamental objects and maybe find discrepancies between the $t\bar{t}H$ signal events and $t\bar{t}+X$ backgrounds. We make an attempt to exploit these with the multivariate method Boosted Decision Trees (BDT), calculated from five physical variables that describe the event. There are two parts of the method - the foundation is a decision tree and adaptive boosting [134] is used to combine a large number of those trees into a single variable. A decision tree is a sequence of rectangular cuts in the space of the input variables, which split this space into multiple hypercubes, allowing to isolate regions with a high concentration of signal or background. At each step the remaining events in a branch are split in two along one of the variables until a limiting case is reached. At this point each final branch will be labelled 'S' or 'B' according to the dominant type. The BDT uses many such trees, but limits the number of consecutive cuts per tree to two or three levels. The collection of trees is called a forest. Each tree in the forest gets a weight according to the fraction of misidentified events, err, during the training, $w = \frac{1-\text{err}}{\text{err}} \in [1, \infty]$. Moreover, the events, which were misclassified in this tree, get re-weighted by multiplying their current weight by w before the next tree is built. This way the distributions of the input variables are changed such that the next tree is forced to focus more on those misclassified events, as they carry more weight. Once the training is complete, the BDT forest has fixed trees with fixed cuts and weights. Each new event goes through each tree and ends up either on a signal (1) or background (-1) final branch, also called a leaf. Moreover, the score is multiplied by the natural logarithm of the tree weight. The cumulative weighted contribution from all trees assigns the event a single number - its BDT score. The higher the score the more likely it is to be a signal event. All steps of the method are described in more detail in [135].

Back to the problem at hand, the five input variables are the invariant mass, transverse momentum and rapidity of the combined $t\bar{t}H$ system as well as the angles of the top and anti-top quarks from the Higgs boson in the $t\bar{t}H$ centre of mass frame. Because of the charged lepton, we are always able to determine which is the top and which the anti-top. To train and later apply the BDT variable we use

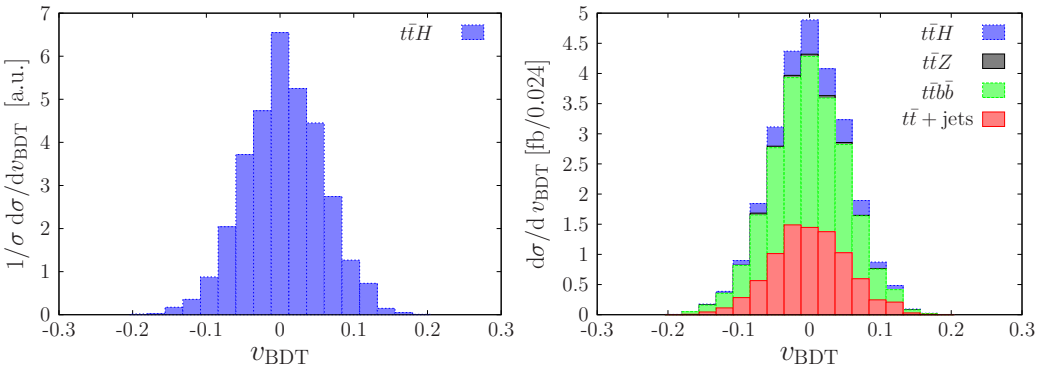


Figure 4.8: Boosted Decision Trees score distribution from 5 variables calculated with the reconstructed $t\bar{t}H$ objects after the mass cut in **T2**. The left figure is the signal $t\bar{t}H$ and the figure to the right contains signal and background.

the TMVA [135] package in the ROOT [136] framework. The forest consist of 850 trees, each with up to three levels. To avoid overfitting, we require that a final leaf be considered only if it contains at least 5% of the weighted signal. Our choice of variables is not exhaustive. The method is limited by the available statistics and the systematic uncertainties of the selected variables' normalisation and shapes. Our implementation is purely as a showcase and neither of the intricacies described above are addressed. The resulting distribution is shown in Fig. 4.8. Even though the results will be described in detail in Sec. 4.4 we point out that the additional analysis steps and techniques designed to improve the S/B ratio of the **T2** selection channel add only modest benefits.

Topologies T3–T5: boosted $\mathbf{t_{had}}$ or boosted H

So far we have covered the phase space of the original analysis in Sec. 4.1 through the two independent selection channels **T1** and **T2**. It is characterised by a boosted hadronic top associated with a fat jet and a boosted Higgs boson associated with another fat jet. Now we relax these conditions and consider two adjacent phase space bins. The first contains events where the HEPTopTagger is unable to identify a single t_{had} candidate, but there is still at least one fat jet in the event. In those events we will be looking to recover a boosted Higgs boson within a fat jet and an unbosted hadronic top from the remaining radiation in the event outside the Higgs

candidate. This bin can be split in the same way as the original analysis into a 1-candidate and 3-candidate fat jets corresponding to selection channels **T3** and **T4**. The last phase space bin, **T5**, includes single fat jet exclusive events, with that jet successfully tagged as a hadronic top. Therefore, we are left to look for an unboosted Higgs boson among the rest of the radiation in the event.

Starting from **T3** and **T4**, we follow the boosted Higgs reconstruction steps in Sec. 4.2.1. We apply a mass drop on the fat jet and continue by grouping the substructures into Higgs candidates. We only keep up to three candidates and treat the 1-candidate and 3-candidate fat jets in independent bins. For each Higgs candidate we recluster all the radiation outside of it into inner and outer jets according to step 6 of the boosted analysis in Sec. 4.1. This time we do not have a reconstructed hadronic top, therefore we require at least a total of four inner and outer jets to match the four quarks from the two top decays. We simultaneously reconstruct the hadronic top and Higgs boson by calculating χ^2 scores for the configurations. Each of them contains a Higgs candidate, a set of three inner or outer jets as a t_{had} candidate, and an additional inner or outer jet as b-quark candidate from the leptonic top. In each configuration there are three permutations among the three hadronic top jets associated with the assignment of the b-quark and W boson. In order to reduce this multiplicity, we always select the assignment with minimum $\Delta m_{\text{W}} = |m_{\text{W}_{\text{reco}}} - m_{\text{W}}|$. Now a configuration is always a unique assignment of final state objects to EW resonance decay products. The χ^2 score is defined in the following way

$$\begin{aligned} \chi^2 &= \chi^2_{\text{top}} + \chi^2_{\text{W}} + \chi^2_{\text{Higgs}}, \\ \chi^2_{\text{top}} &= \frac{(m_{\text{t}_{\text{had}},\text{reco}} - m_{\text{t}_{\text{had}},\text{max}})^2}{\sigma^2_{\text{t}_{\text{had}}}}, \\ \chi^2_{\text{W}} &= \frac{(m_{\text{W}_{\text{had}},\text{reco}} - m_{\text{W}_{\text{had}},\text{max}})^2}{\sigma^2_{\text{W}_{\text{had}}}}. \end{aligned} \tag{4.2.4}$$

The χ^2_{Higgs} and χ^2_{top} are identical to the ones defined in Sec. 4.2.1 with the exception that χ^2_{top} involves the hadronic top reconstructed mass $m_{\text{t}_{\text{had}},\text{reco}}$. The parameters in χ^2_{W} are extracted from fitting a Gaussian distribution to the histogram of the reconstructed W boson mass by the top tagger in [123] in the pure A_1 topology, which is displayed in the right plot of Fig. 4.3. Similarly to the doubly boosted

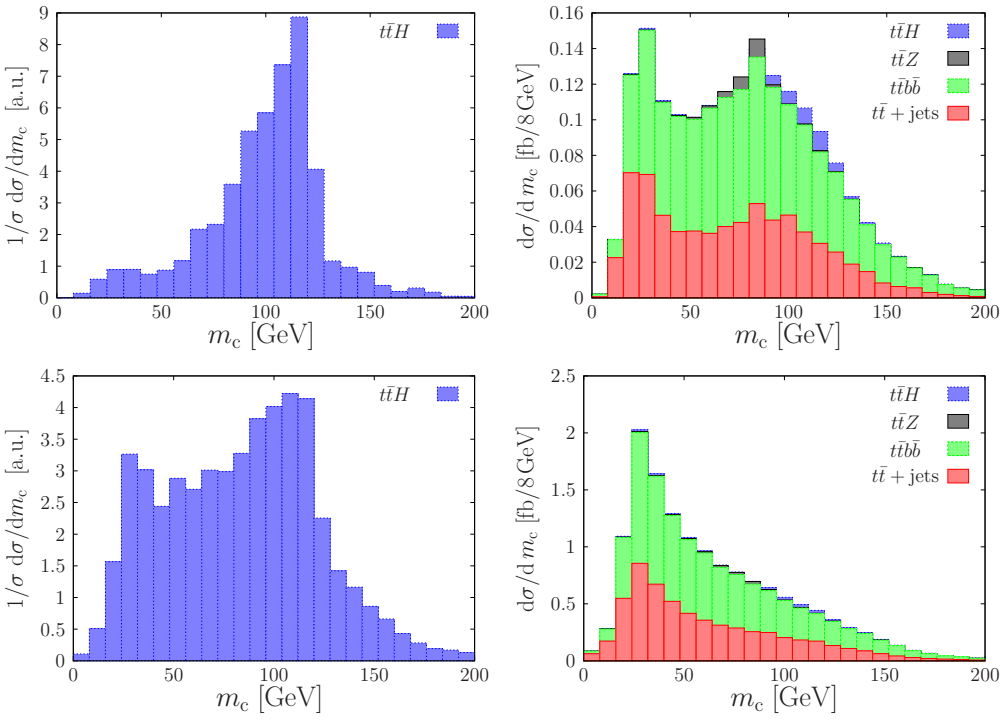


Figure 4.9: m_c distribution obtained from the selection channels without any top tags - **T3** (top) and **T4** (bottom). The left figures show the $t\bar{t}H$ signal only and the figures to the right contain signal and background.

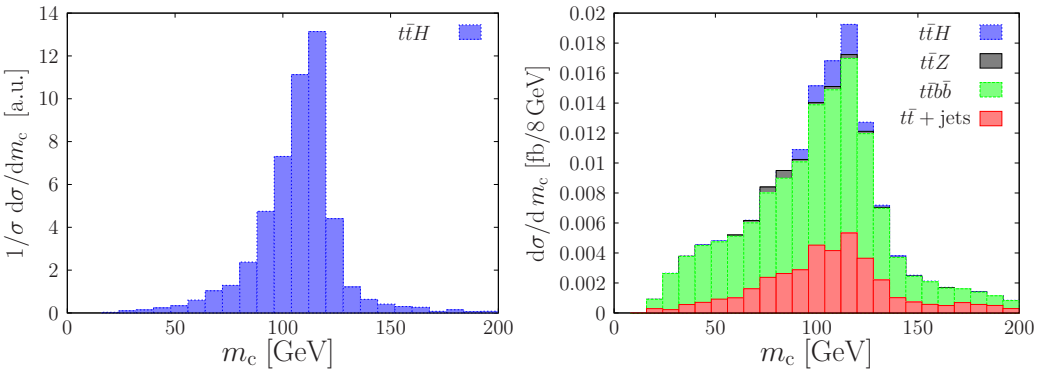


Figure 4.10: m_c distribution obtained from the selection channel with only one fat jet that has been top-tagged (**T5**). The left figure is the signal $t\bar{t}H$ and the figure to the right contains signal and background.

analysis of channel **T2**, the configurations are ordered by χ^2 in ascending order and the lowest 25% are kept. To be consistent with the rest of the analysis, there are only three b-tag requirements on each remaining configuration: two within the Higgs candidate filtered subjets and one for the designated t_{lep} b-quark candidate. One could also require that the hadronic top b-quark candidate gets a tag. The Higgs candidate mass of all surviving configurations is recorded in the histograms in Fig. 4.9. As expected, we see that the 1-candidate channel **T3** recovers a much cleaner peak than the 3-candidate channel **T4**, which also leads to a better S/B ratio. A benefit of both these selection channels is that the number of remaining $t\bar{t}H$ events is an order of magnitude larger then the doubly boosted channels **T1** and **T2**.

The last channel, **T5**, is the one with a boosted top but no more fat jets to recover a boosted Higgs. Therefore, we recluster all hadrons outside the fat jet into C/A $R = 0.4$ jets with $p_T > 30$ GeV and require exactly 3 of them to get a b-tag. What remains is to reconstruct a Higgs and a leptonic top. This was already done for **T2** in Sec. 4.2.1. We use the same χ^2 score defined in Eq. (4.2.2) and evaluated on Ht_{lep} configurations, each containing two b-jets as a Higgs candidate, the remaining b-jet, the isolated lepton and one of two reconstructed neutrinos. This time the number of configurations is limited to six by construction (three options for the leptonic b-quark and two options for the neutrino p_z). Another difference is

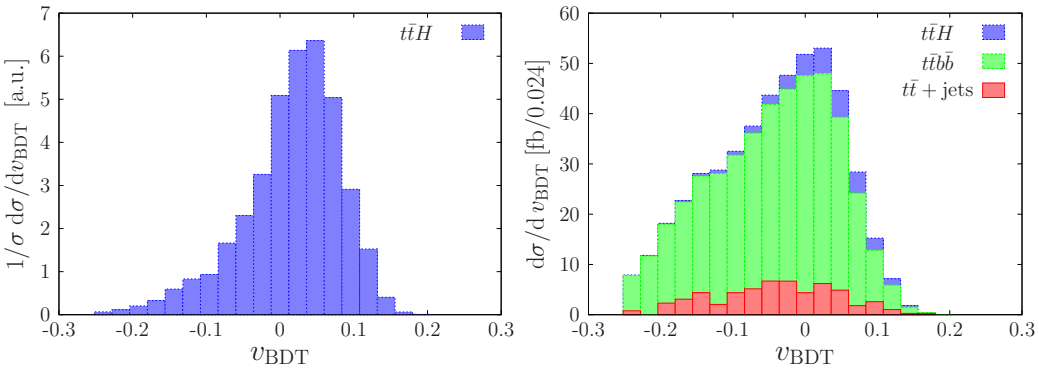


Figure 4.11: Boosted Decision Trees score distribution from 7 variables calculated from objects in the non-boosted analysis. The left figure is the signal $t\bar{t}H$ and the figure to the right contains signal and background.

that we only take the configuration with the best (smallest) value of χ^2. The mass distribution of the selected Higgs candidate is presented in Fig. 4.10. Even though the S/B ratio of this channel is comparable to **T3**, it has five times smaller signal yield and the background distribution is severely biased.

4.2.2 MVA Without Boost

In the previous sections 4.1-4.2.1 we intentionally split the $t\bar{t}H$ phase space into regions with different boosted massive particles in an attempt to improve the S/B ratio in this important but difficult Higgs production channel. In this section we go back to treating the entire phase space in the same way and see if the boosted analysis has benefits over a more classical approach. First we define the physical objects that are going to be used. We require a single isolated lepton and cluster the hadrons into C/A R=0.4 jets with $p_T > 30$ GeV. The events we are interested in have at least six such jets and in addition exactly four of them must be b-tagged. At this point we keep all b-jets and the two hardest non-b-jets for a total of six.

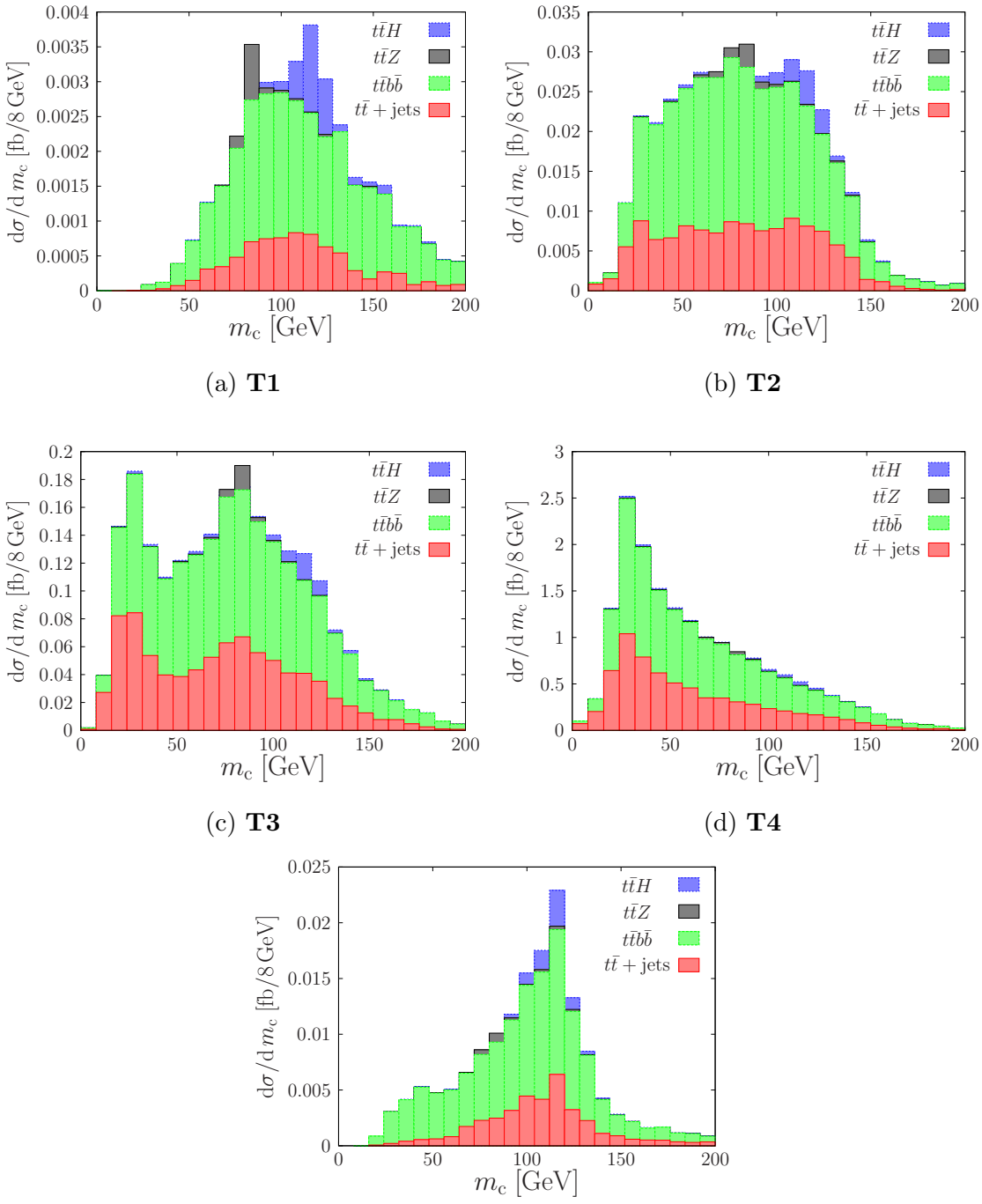
These six jets, which will be called unambiguously $(\mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3, \mathrm{b}_4, \mathrm{q}_1, \mathrm{q}_2)$, in combination with the isolated lepton and the missing transverse momentum $\not{E}_T$ are used to evaluate simple kinematic variables that will be combined in a MVA. The numbering scheme in the jet name designation indicates the descending order in p_T. The variables in question are: $\Delta m_{\mathrm{H}} = \min_{\mathrm{ij}} \left| m_{\mathrm{H,max}} - m_{\mathrm{b_i b_j}} \right|$, $p_{\mathrm{Tq_2}}/p_{\mathrm{Tq_1}}$, $\max_{\mathrm{ij}} \Delta R_{\mathrm{b_i b_j}}$,

$\min_i \Delta R_{W,b_i}$, $\Delta\phi_{\not{E}_{\mathbf{T}},\mathbf{b_3}}$, $\Delta R_{\ell,\mathrm{b}_3}$, $\Delta R_{\mathrm{W,b}_4}$. It is obvious that a multitude of variables of this type can be constructed if we move around the possible input particles. However, these seven get the highest rank, as defined in [135], when we construct a BDT with all possible permutations. Therefore, we define a new BDT with these seven variables only and use it to separate $t\bar{t}H$ from $t\bar{t}+X$ and compare to the boosted analysis. The resulting distributions are shown in Fig. 4.11. Despite the undeniable practical benefit in the event classification provided by the BDT, it is not obvious how to interpret the physics behind the resulting variable. Nevertheless, it seems to provide comparable results to our boosted methods, with some caveats, which are described in more details in Sec. 4.4.2.

4.3 Effects from b-jet energy correction

Throughout the chapter we have only been noting the effects of the neutrino decay products from B-meson decays on the mass distributions of the EW resonances. Specifically, the Higgs peak, which is formed from two b-quarks, suffers severely from the missing momentum of the B-meson neutrinos as it is shifted and its low mass tail elongated towards the background-rich part of the mass spectrum. The experimental groups in ATLAS and CMS have developed energy-correction techniques that account for the loss of energy in b-jets. However, this analysis is not sophisticated enough to make use of such techniques. In particular we do not perform any detector simulations. Despite that, we are in a position to show the extreme cases of such energy correction. So far we have displayed what happens if these effects are completely ignored. But it is easy to also show what a perfect B-meson energy reconstruction would yield by including the neutrinos into the mix of final state visible particles. In the results section we also present the sensitivity to $t\bar{t}H$ events in this most optimistic outcome.

The positive effect from the neutrino inclusion on the Higgs candidate mass distribution in each of the five selection channels **T1** to **T5** is presented in Fig. 4.12. This effect is most obvious and beneficial for **T1**. The Higgs peak sharpens and the S/B ratio increases to 40% with the same signal yield as long as the mass



(a) **T1** (b) **T2**

(c) **T3** (d) **T4**

(e) **T5**

Figure 4.12: Distributions in the Higgs-candidate mass m_{c} after three b-tags for the various selection topologies as in Figs. 4.6–4.10, but including neutrinos in the reconstructed B-hadrons.

window is adjusted to account for the peak shift to a larger mass. Another obvious characteristic is the sharp Z boson peak. Together with the signal depleted regions, it can be used in a data-driven analysis to estimate the signal strength of the Higgs peak and to confine the background continuum uncertainty. Unfortunately, such obvious benefits are not present in the more frequent, but less pure, channels **T2** to **T5**.

4.4 Results from the $t\bar{t}H$ selection strategies

In this section the results from the various analyses and selection channels are reported in two ways. One is as S/B ratios after cuts that select regions in the distribution with high signal concentration. The signals S and background B refer to the Standard Model expectation of the number of events of type $t\bar{t}H(b\bar{b})$ and of type $t\bar{t}+X$ respectively, where X stands for $b\bar{b}$, light jets, and Z. Since we are interested how sensitive the $t\bar{t}H$ search strategies are to deviations from the total SM expectation ($S+B$ in the language of this paragraph), the results are also displayed in the form of 95% CL limits on the signal strength μ at different integrated luminosities. Here μ is defined in such a way that a measurement consistent with the SM yields $\mu = 0$, while positive and negative contributions from BSM physics or deviations of the expected SM result in $\mu > 0$ and $\mu < 0$ respectively. Thus $\mu = \frac{\sigma^{\text{obs}} - \sigma^{\text{SM}}_{\text{S+B}}}{\sigma^{\text{SM}}_{\text{S}}}$. The signal strength is normalised to the SM cross section of the semileptonic $t\bar{t}H(b\bar{b})$.

All of our search strategies lead to one of two types of final distributions. Usually, this is a Higgs candidate mass distribution m_{c}, but it could also be a MVA score v_{BDT}. These are the distributions from which the CL limits are extracted with a two-sided frequentist test using the profile likelihood test statistic and CL_s to quote the confidence level. To calculate the statistical model and build the profile likelihood distributions for different values of μ we use the RooStats framework [137]. The null hypothesis corresponds to $\mu = 0$, the expected number of events according to the SM, and we vary μ in the alternative hypothesis to find the upper and lower limits on the signal strength from BSM contributions that this analysis can impose. The

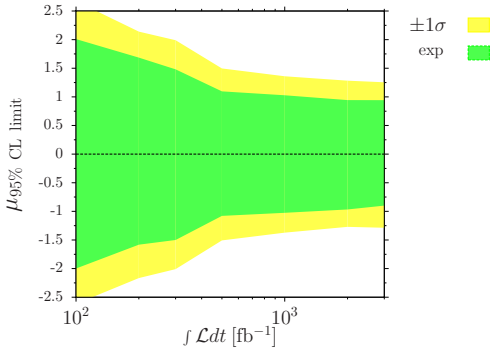
stage	$t\bar{t}H$	$t\bar{t}b\bar{b}$	$t\bar{t}$+jets	$t\bar{t}Z$	S/B
MC level	94	7.3×10^3	2.6×10^5	50	3.5×10^{-4}
1 lepton	60	4.7×10^3	1.6×10^5	22	3.6×10^{-4}
>1 fat jets	15	400	9.5×10^3	5.9	1.5×10^{-3}
1 top tag	4.8	110	2.6×10^3	1.9	1.8×10^{-3}
3 b-tags	0.59	7.6	4.2	0.25	0.049
m_c cut	0.2	0.9	0.48	0.023	0.14

Table 4.3: Signal and background cross sections in femtobarn and S/B ratios at different stages of the boosted analysis of Section 4.1.

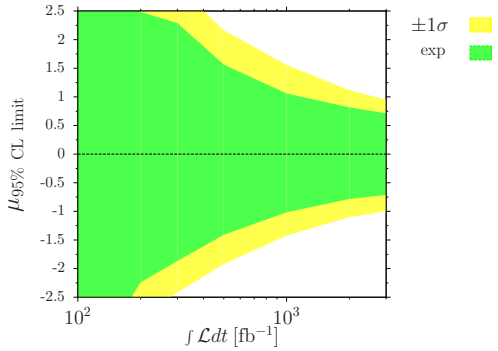
background uncertainty is treated in two ways. Either as a Gaussian centred at the null hypothesis expectation value with a flat 15% standard deviation (Fig. 4.13) or in a more optimistic scenario a decreasing standard deviation as the square root of the integrated luminosity above $\int \mathcal{L}dt = 300\,\text{fb}^{-1}$ (Fig. 4.14). The bands in those figures cover the μ values that are too close to the background hypothesis to be excluded at 95% CL with the analysis and this choice of error. The green band assumes that the observed value is the median of the null hypothesis, while the yellow band assumes a 1σ deviation from the median. Appendix A elaborates in more detail the statistical methodology.

4.4.1 Standard boosted analysis

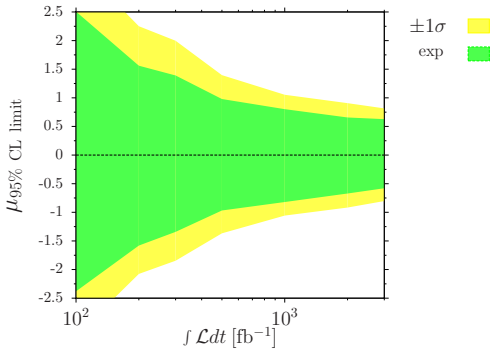
The most notable conclusion from the results of the standard boosted analysis in Sec. 4.1 is that the sensitivity to $t\bar{t}H(b\bar{b})$ events is worse compared to the very similar analysis in [123]. Since the latter was proposed before the Higgs mass was known, we can extrapolate the S/B ratio between the two closest mass points $m_H = 130$, 120 GeV. The 2009 analysis found for these choices of the Higgs mass the ratios $S/B = 42\%$, 28% respectively. Compared to that, the analysis in section 4.1 obtained $S/B = 14\%$, which is significantly less. The cross section after the different steps as well as the corresponding S/B ratio are presented in Table 4.3. We explain the discrepancy with the relative corrections to the signal and back-



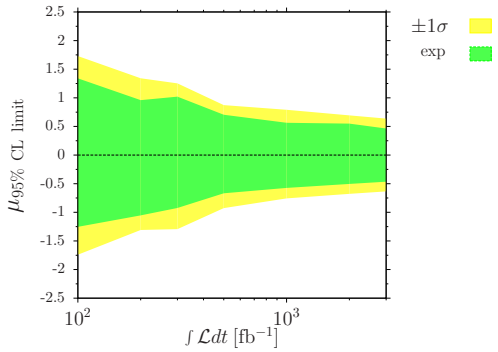
(a) Analysis of Sec. 4.1 including all relevant topologies (**T1** and **T2**).



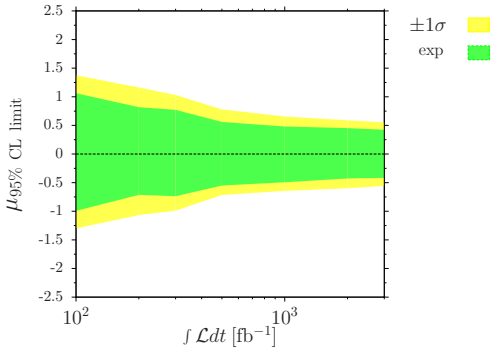
(b) Analysis of Sec. 4.2 limited to topology **T1**.



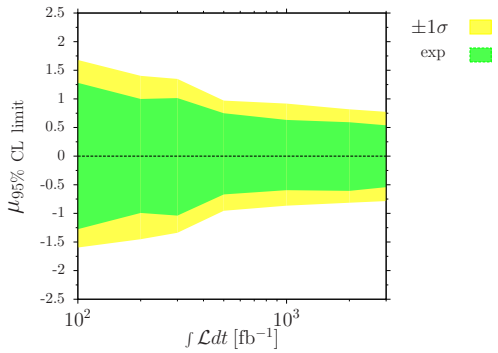
(c) Analysis of Sec. 4.2 including topologies **T1** and **T2**.



(d) Analysis of Sec. 4.2 including all topologies (**T1**–**T5**).

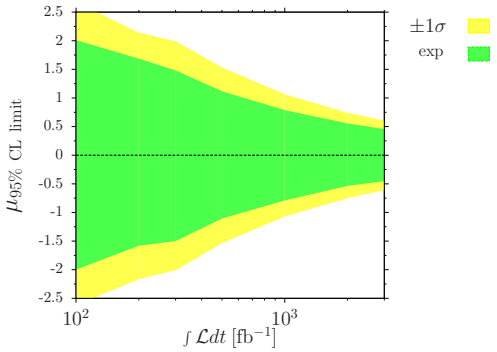


(e) Analysis of Sec. 4.2 including all topologies (**T1**–**T5**) and neutrinos in B-decay reconstruction.

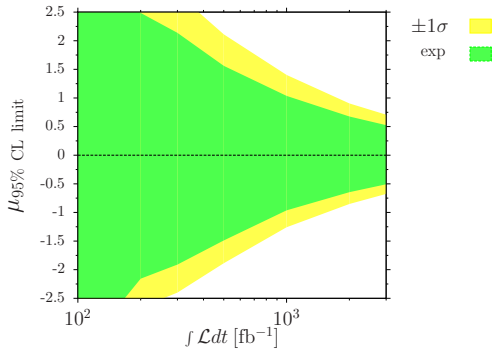


(f) Unboosted BDT analysis of Sec. 4.2.2.

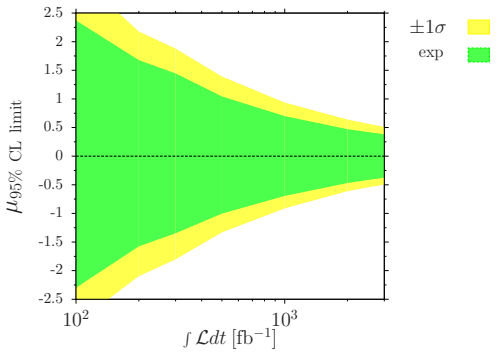
Figure 4.13: Two-sided 95% CL limit of the signal strength μ as a function of the integrated luminosity assuming a constant 15% normalisation uncertainty for the SM background.



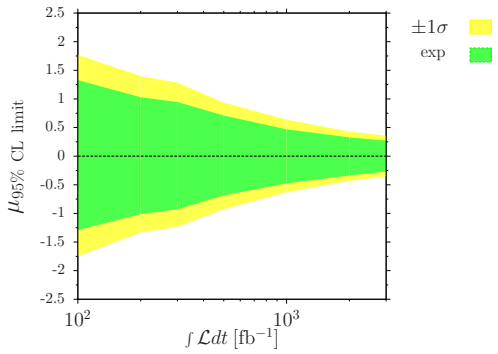
(a) Analysis of Sec. 4.1 including all relevant topologies (**T1** and **T2**).



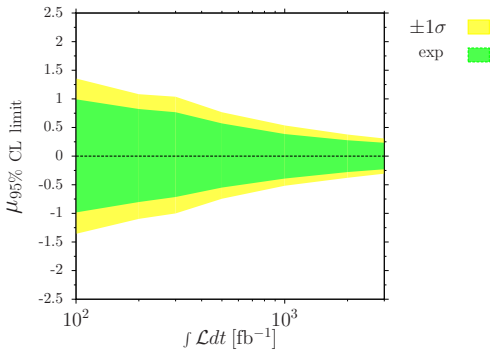
(b) Analysis of Sec. 4.2 limited to topology **T1**.



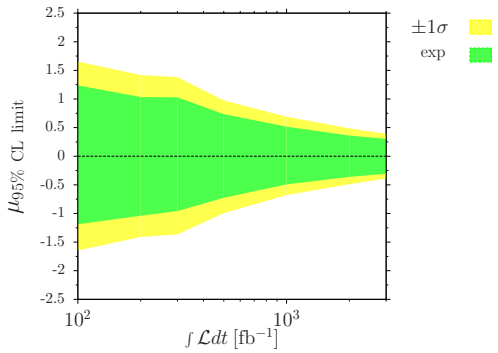
(c) Analysis of Sec. 4.2 including topologies **T1** and **T2**.



(d) Analysis of Sec. 4.2 including all topologies (**T1**–**T5**).



(e) Analysis of Sec. 4.2 including all topologies (**T1**–**T5**) and neutrinos in B-decay reconstruction.



(f) Unboosted BDT analysis of Sec. 4.2.2.

Figure 4.14: Two-sided 95% CL limit of the signal strength μ as a function of the integrated luminosity assuming a normalisation uncertainty for the SM background that remains constant at 15% level up to 300 fb^{-1} and scales as $1/\sqrt{\int \mathcal{L}dt}$ for higher integrated luminosities.

ground simulations, the change in b-quark tagging methodology and the switch to HEPTopTagger.

The most detrimental effect is the overall increase by 35% of the total background, mainly accounted for by a huge contribution from $t\bar{t}+$ jets compared to [123], and a 30% drop in the $t\bar{t}H$ cross section that remains after the final cut. Both the signal and background simulations are much more reliable in the current analysis, as they rely on NLO accuracy tools as described in Sec.II of [3] compared to $LO + PS$ in [123]. The correction is especially large for $t\bar{t}+$ light jets events, which were generated in [123] from a $t\bar{t} + 1$ jet LO matrix element. For this study the events were generated from an inclusive $t\bar{t}$ matrix element accurate up to 2 light jets at LO and normalised to NLO cross section.

There is an additional contribution to the relative increase in $t\bar{t}+$ jets events in the final selection cut from the b-tagging. There are two effects at play - one acts to reduce the signal while the other boosts the background. Looking at the effect of triple b-tagging on the $t\bar{t}H$ sample (Table 4.3 rows 4 and 5), the signal is reduced by a large factor of 8. A naive application of the incorporated b-tag efficiency $\epsilon_b = 0.7$ suggests that the signal should be suppressed only by a factor of $\epsilon_b^{-3} \approx 3$. There must be another source of efficiency loss within the b-tagging method. This comes from an inherent mismatch between geometrically defined objects, like jets and subjets, and the particles that initiate them. We only apply the ϵ_b tagging efficiency to jets and subjets that contain a B-meson within their radius. Especially when it comes to the $R = 0.2$ subjets of the Higgs candidate, the probability that the B-meson is away from the jet, initiated by the b-quark, is small but significant. Even a single mismatch reduces the efficiency of the event by a factor of 70. Therefore, in order to preserve the naive expectation, not a single mismatch must occur. This is where the difference between $t\bar{t}H$ and $t\bar{t}$ + jets lies. Since there are three b-quarks in the former, compared to only one in the latter, the probability of no mismatch is smaller for the signal. Therefore, the signal will be reduced more often by this artefact of our b-tagging method.

Looking at the b-tagging suppression factor of the $t\bar{t}$ + jets background we see that three tags reduce its contribution by $2600/4.2 \simeq 620$. However, given the tiny

mistag rate $\epsilon_{mt} = 0.01$, one would expect a stronger suppression. Naively, for example, $n = 4$ light jets and a single b-jet give a suppression factor $[n(n-1)/2\,\epsilon_{mt}^2\,\epsilon_b]^{-1} \simeq 2400$. There is an event topology, where the hadronic top is reconstructed without a b-quark, which leaves both b-quarks for Higgs and t_{lep} tagging. Then the configurations with maximum contribution involve both b-jets and the suppression factor becomes $[f_T\,n\,\epsilon_{mt}\,\epsilon_b^2]^{-1}$. This factor will depend upon the contribution of this topology. However it will dominate even for $f_T \simeq 0.05$ and give a suppression factor of around 1000, which is more comparable to the observed value. A jet topology study similar to the one in Sec. 4.1.1 is needed to determine the exact f_T fraction. In any case the arguments above show that the relative suppression of signal and background events from b-tagging is very sensitive to the exact method of b-tagging, top-tagging, and the correct simulation of multi-jet emissions in the $t\bar{t}+X$ samples. There is still the possibility of including a fourth b-tag to reduce the $t\bar{t}$ + light jets sample. It will improve relative signal and background contributions, but not by a naive factor of 70.

We mentioned that the HEPTopTagger also contributes to the different results in Sec. 4.1 and [123]. The jet topology study in Sec. 4.1.1 showed that the HEPTopTagger is very consistent in its top reconstruction even when additional hard radiation pollutes the fat jet candidate. The top tagging method in [123] has a different behaviour. It has a larger efficiency (and mistag rate), which increase further as more radiation is added to the top decay products. Therefore, it is a less effective tool in removing non-$t\bar{t}$ types of backgrounds. However, since we do not consider these, the effect of the old tagger is to increase the overall event count.

The sensitivity of the analysis from Sec. 4.1 is presented in Fig. 4.13a-4.14a. We can see that, with a constant systematic uncertainty of 15%, it is not sensitive to $|\mu| \lesssim 1$ and, therefore, the only BSM contributions that can be excluded must be larger (in absolute terms) than the Standard Model cross section for $t\bar{t}H$. Moreover, this sensitivity is largely constant for most integrated luminosities over Runs 2 and 3, meaning that the exclusion is limited by the systematic uncertainty. A more generous assumption, reducing this uncertainty as $1/\sqrt{\int \mathcal{L}dt}$, will allow 95% CL exclusions of models that predict $|\mu| \gtrsim 0.5$ at the final integrated luminosity $\int \mathcal{L}dt =$

3000 fb^{-1}.

4.4.2 Improved boosted analyses T1–T5 and unboosted MVA approach

It is interesting to see how much the extension of the standard boosted analysis improves the sensitivity to BSM contributions, and also how this sensitivity compares with an unboosted analysis. The results from the different selection channels from Sec. 4.2, including the unboosted one, are summarised in Table 4.4. Isolating the 2-prong Higgs fat jets into a category of their own, **T1**, drastically improves the Higgs mass peak sharpness. After selecting the events in the chosen m_{c} window, S/B jumps from 14% to 23%. However, the improvement in the purity comes at a price. We already discussed that the **T1** channel accounts for about 20% of the total signal that contributes to the standard boosted analysis. The rest falls within the **T2** bin. The addition of the ellipticity cut $\hat{t} < 0.2$ improves the S/B to 27% and costs only 15% of the signal yield. Even at 300 fb^{-1}, the target integrated luminosity at the end of Run 2, we expect to detect 12 $t\bar{t}H(b\bar{b})$ events with topology **T1**. Nevertheless, the improved S/B ratio will outweigh the low statistics drawback as the LHC moves into the high luminosity run. At 3000 fb^{-1} and a systematic uncertainty of 15%, the **T1** channel on its own will be able to exclude $|\mu| > 0.7$, which is an improvement on the standard boosted analysis (see Fig. 4.13). There are four more statistically independent channels **T2**-**T5**, which can be added in the statistical analysis and improve the exclusion limit.

Despite the multiple efforts to improve the S/B ratio of the events in the complementary **T2** channel with 3 Higgs candidate jets, only a cut on v_{BDT}, defined by combining five variables associated with the reconstructed $t\bar{t}H$ system, is able to bring the ratio from 13% to 15%. But this small boost is accompanied by a sizeable drop in cross section (a factor of three). All other attempts at exploiting different physics arguments - a cut on χ^2, the colour flow sensitive jet shape $\hat{t}$ and helicity angle between the lepton and the third b-quark - do not bring the S/B ratio up. Still, using the **T2** channel without the modifications and **T1** in a common profile likelihood, yields a drop in the exclusion limit from $|\mu| \simeq 0.7$ to $|\mu| \simeq 0.6$. The three

Analysis	stage	$t\bar{t}H$	$t\bar{t}b\bar{b}$	$t\bar{t}$+jets	$t\bar{t}Z$	S/B
T1	before b-tag	1.1	27	690	0.43	1.5×10^{-3}
	3 b-tags	0.075	0.77	0.37	0.032	0.064
	m_{c} cut	0.042	0.13	0.053	2.0×10^{-3}	0.23
	$\hat{t}$ cut	0.035	0.089	0.038	9.5×10^{-4}	0.27
T2	before b-tag	12	240	4.6×10^{3}	4.5	2.5×10^{-3}
	3 b-tags	0.25	3.0	1.5	0.11	0.054
	m_{c} cut	0.14	0.66	0.36	0.01	0.13
	v_{BDT} cut	0.044	0.18	0.1	0.0031	0.15
T3	before b-tag	51	1.2×10^{3}	1.9×10^{4}	18	3.0×10^{-3}
	3 b-tags	1.0	17	11	0.48	0.04
	m_{c} cut	0.53	3.2	2.0	0.032	0.1
T4	before b-tag	630	1.5×10^{4}	2.2×10^{5}	210	3.0×10^{-3}
	3 b-tags	5.6	130	92	2.2	0.02
	m_{c} cut	1.5	16	10	0.2	0.06
T5	before b-tag	4.2	220	5.7×10^{3}	1.5	7×10^{-4}
	3 b-tags	0.14	1.6	0.65	0.036	0.06
	m_{c} cut	0.094	0.6	0.28	0.011	0.11
MVA	>5 jets	14	420	6.0×10^{3}	5.1	2.2×10^{-3}
	4 b-jets	1.5	19	2.9	0.52	0.066
	v_{BDT} cut	0.041	0.16	0.033	2.4×10^{-3}	0.21

Table 4.4: Signal and background cross sections in femtobarn and S/B ratios at different stages of the various boosted analyses (**T1**–**T5**) of Section 4.2.1 and for the unboosted MVA analysis of Section 4.2.2.

bins with only a single boosted object (**T3** to **T5**) have comparable or even smaller S/B ratio to the **T2** channel, ranging between 6% and 11%. Yet, they do not suffer from low statistics. When combined with **T1** and **T2**, they reduce the limit further to $|\mu| = 0.45$.

An alternative approach (but not statistically independent from **T1** through **T5**) is the unboosted MVA analysis presented in Sec. 4.2.2. It is done in the spirit of the ATLAS and CMS collaborations' analyses for the Run 1 data, even though it lacks the comprehensive approach and detail. The main purpose is to compare how the results of this classic strategy compares with the boosted search. The cut on the BDT score, that leads to the result in the last row of Table 4.4, is rather stringent. This is because to fairly compare the S/B ratio of this method to **T1**, we should allow the cut to leave comparable signal yield. With that in mind, the S/B ratio of the unboosted MVA analysis is very close but slightly under the **T1** ratio at 21% compared to 23%. Yet this is not the optimal cut in terms of statistical significance. With a looser constraint on v_{BDT} it is possible to increase the signal yield almost by an order of magnitude while dropping the S/B ratio to 18%. If we use the BDT distribution to define a profile likelihood, at $3000\,\mathrm{fb}^{-1}$ it is possible to exclude $|\mu| \gtrsim 0.55$. Therefore, the unboosted MVA analysis fairs better than the **T1** channel on its own, $|\mu| \simeq 0.7$, but not as well as the combined sensitivity of all five boosted channels $|\mu| \simeq 0.45$. There are some benefits of the **T1** channel over the MVA analysis, which have not been employed here but could definitely tilt the balance. There is a sharp Higgs resonance peak over a much more slowly varying background, which can be used to constrain the background uncertainty by analysing signal-depleted regions in the distribution. Moreover, if the b-jet energy correction allows for it, the $Z \to b\bar{b}$ peak can become very distinct from the background, which would provide additional avenues to constrain the uncertainty in a data driven way.

So far the data has only been interpreted under the assumption that the systematic uncertainty remains constant (Fig. 4.13). Yet, the final integrated luminosity of $3\,\mathrm{ab}^{-1}$ will not be reached for another decade. In this time, new advances may lower the theoretical uncertainty. Moreover, the possibility of data driven constraints on the nuisance parameters means that treating the uncertainty as a constant may not

be optimal. In Fig. 4.14 we treat the systematic uncertainty in a similar way to a statistical uncertainty in that we make the error proportional to the inverse square root of the integrated luminosity beyond $300\,\mathrm{fb}^{-1}$. This change shifts the balance between the importance of signal purity and signal yield at large luminosity. Whereas before the S/B ratio played the most important role in determining the sensitivity of the analysis at $3000\,fb^{-1}$, now the final signal count becomes vital. Therefore, the unboosted analysis, having looser kinematic constraints, outperforms each individual boosted channel. For example **T1** now excludes any $|\mu|$ larger than 0.5, but the unboosted MVA goes as low as 0.29. Nevertheless, the combination of **T1** to **T5** provides the best exclusion at $|\mu| \gtrsim 0.26$. In an even more optimistic scenario where the b-jet energy correction works nearly perfectly, the combined boosted analysis can be sensitive to changes in the $t\bar{t}H$ cross section as low as 20% of the Standard Model expectation value.

4.5 Summary of $t\bar{t}H$ tagging

This chapter was dedicated to evaluating the LHC capacity to measure the signal strength μ of the semileptonic $t\bar{t}H$ Higgs production channel with the Higgs decaying hadronically $H \to b\bar{b}$. Even though a similar study [123], intended for Higgs discovery through this mode, showed very promising sensitivity to deviations from the SM expected value, through a combination of improved event simulation and particle selection we have been able to acquire only loose exclusion limits on μ. In order to improve the result, we have split the phase space of the $t\bar{t}H$ event that is already analysed with boosted techniques and extended the search to other phase space regions with only a single boosted object. The combined contribution of all independent regions leads to a noticeable improvement of the expected 95% CL exclusion limit. With those changes and under optimistic assumptions about the theoretical and experimental uncertainties, the limit may shrink to $|\mu| = 0.2$.

Chapter 5

Conclusions

The LHC already proved to be a successful endeavour with the discovery of the missing link in the Standard Model, the Higgs boson. There are still many questions that need to be answered though. Some of them are straightforward to define - like the properties of the Higgs boson. Other questions about what lies beyond the SM are more open ended. In any case, the energy frontier requires new search techniques to isolate interesting signal from overwhelming background processes. Particularly, as the collisions are hadronic, QCD-infested events need to be well understood and the important underlying event structure reconstructed. This thesis described the attempts to ameliorate the signal extraction for three different processes.

In Chapter 2 we proposed a novel use of the shower deconstruction method in the long-sought-after goal of tagging quark-initiated jets from gluon-initiated jets. Even though we did not find a smoking-gun type of indicator, we were able to improve the quark-to-gluon ratio above the performance of other taggers in a broad range of useful kinematic regimes. Moreover, we learned that the distributions of variables based on gluon and quark jets are never symmetric between the two. Thus, the quark tagging has always been better than gluon tagging. This seems to be a property of the gluon and quark evolutions as opposed to an artefact from the choice of tagging variable because it is a persistent observation.

The next chapter (3) was dedicated to very boosted jets at high virtuality, where the emission of a heavy electroweak boson is modified by logarithms. We proposed techniques to isolate a W within a larger fat jet and quantified to what devia-

tions from the expected rate our search strategies are sensitive. We looked for both hadronically and leptonically decaying W bosons. Unsurprisingly, the leptonic channel is purer with respect to QCD-only background and is the easier option at the energies of the LHC. However, at very large boosts, which might be achieved in a 100 TeV machine, the isolation criterion may hinder the leptonic channel to an extent that hadronic substructure techniques would be the better choice for collinear W identification.

Finally, Chapter 4 developed an analysis for a very inclusive measurement of the signal strength in the semileptonic $t\bar{t}H(b\bar{b})$ process in the second and third runs of the LHC. We looked at different scenarios where the hadronic top or the Higgs or both are modestly boosted. Thus, we were able to apply substructure techniques and constrain the combinatorial background of this busy event, while preserving as much signal as possible. We also compared our results to a more standard multivariate analysis without boost requirements. Even though our new approach provides a marginally better sensitivity in terms of S/B ratio in the signal-rich part of the phase space, it allows for much better data-driven background estimation by providing a known peak structure (the Z resonance) and a Higgs peak structure on top of $t\bar{t}b\bar{b}$ background.

Appendix A

Statistical Method

To evaluate the sensitivity of the analysis techniques of chapters 3 and 4, we need to model the probability of the final results under different hypotheses. Once a statistical model is provided for each independent result, we can combine them through a single test statistic, evaluate the distribution of the statistic under different hypotheses, and provide exclusion limits and p-values.

In both searches, collinear W bosons and $t\bar{t}H$ events, the same building blocks are used to form the full statistical model of the analysis results. The steps of our algorithms lead either to a distribution of a variable, such as a mass distribution or jet shape distribution, in the form of a histogram, or event counts in a selection window, potentially one from a series of independent selection channel. From a statistical point of view, there is no distinction. Each bin is a counting experiment. Therefore, the most suitable probability mass function of the experimental results is a Poisson distribution with mean given by the sum of the expected background events b and the expected deviation from it by the new hypothesis s. In a counting experiment the number of hits is

$$P_c(n|s,b) = \frac{(s+b)^n e^{-(s+b)}}{n!}\,. \tag{A.0.1}$$

Moreover, as long as the different bins represent independent regions of the phase space, the probability of getting a particular result n_i in bin i and result m_j in bin j is $P(n_i \cap m_j) = P(n_i)P(m_j)$. This is always true for the different histogram bins of a single variable because a single collision event cannot fall into more than one bin.

When different selection channels are involved, this is not always the case. However, in our $t\bar{t}H$ search we have constructed the selection channels in such a way that they are statistically independent and the simple multiplication of probabilities applies.

So far the statistical model of the results from our physical analyses covers the possibility of statistical fluctuations, but there are systematic uncertainties associated with the parameters of this statistical model. They could come from a multitude of sources, such as the theoretical accuracy of the expected cross sections for different models, the accuracy of the selection efficiencies of the analysis methods, the integrated luminosity of the experiment, the experimental resolution, the efficiency of the trigger conditions and many more. For example, the search for the Higgs boson [138] incorporated ≈ 200 such uncertainties. An accurate modelling of this magnitude is beyond our analyses, but we do wish to show the limitations of the proposed methods given a systematic uncertainty of an ad hoc chosen proportion on the parameters of the counting model.

The equation for the counting model, Eq. A.0.1, contains only two parameters (we have combined all efficiencies, cross section and luminosity information into expected number of counts s and b). One of them s is the parameter that determines the theoretical model we wish to test. We assume no information about this parameter and we wish to estimate or constrain it from the experimental results. The parameter b contains our current knowledge, but this knowledge is not absolute, so we apply a systematic uncertainty in the form of a Gaussian distribution around the estimate of b from Monte Carlo results $\mu = b$ and with a standard deviation as a selected proportion of b, $\sigma = \epsilon \cdot \mu$,

$$P_s(b'|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b'-\mu)^2}{2\sigma^2}} . \tag{A.0.2}$$

The way this error is implemented in the statistical model is different in Chapters 3 and 4, but the common feature it that it sets a limit to which collecting more data improves the accuracy of the parameter estimation. This effect is most easily seen in Fig.3.7, where the improvement in the sensitivity saturates at large luminosity. In the next two subsections, we describe exactly how we build the statistical models, the test statistic and how we extract limits and p-values for the results in chapters 3 and 4 respectively.

A.1 W boson tagging

The statistical analysis of the results in Chapter 3 closely follow the modified frequentist approach [118] used in LEP. We model the outcomes in our collinear W emission search by considering the number of hits in each bin in a mass or jet shape distribution from Sec. 3.1 as a Poisson random variable. Under the assumption that the background is exactly known, the results are described by

$$\prod_i P_c(n_i|s_i, b_i). \tag{A.1.3}$$

In the context of the W splitting enhancement factor $f_W \equiv f$ from Eq. 3.0.2, the null and alternative hypotheses are $H_0 = f_0 \boldsymbol{s}_{\mathrm{SM}} + \boldsymbol{b}_{\mathrm{SM}}$ and $H_\mathrm{a} = f_1 \boldsymbol{s}_{\mathrm{SM}} + \boldsymbol{b}_{\mathrm{SM}}$, where $\boldsymbol{s}_{\mathrm{SM}} = (s_{\mathrm{SM\,1}}, s_{\mathrm{SM\,2}}, ...)$ are the SM expectation values for the contribution from W emissions to each bin in the histogram according to the event generator. Analogously, $\boldsymbol{b}_{\mathrm{SM}} = (b_{\mathrm{SM\,1}}, b_{\mathrm{SM\,2}}, ...)$ is the contribution from the QCD background. According to the language of Eq. A.1.3, the parameters s and b for each bin i correspond to $s_i = (f_1 - f_0)s_{\mathrm{SM\,i}} + b_{\mathrm{SM\,i}} = \Delta f s_{\mathrm{SM\,i}} + b_{\mathrm{SM\,i}}$ and $b_i = f_0 s_{\mathrm{SM\,i}} + b_{\mathrm{SM\,i}}$ respectively. The test statistic that combines information from all bins is the likelihood ratio:

$$q_W(\boldsymbol{n}) = \prod_i \frac{P_c(n_i|H_a)}{P_c(n_i|H_0)} = \prod_i \frac{(s_i + b_i)^{n_i}\, e^{-(s_i+b_i)}}{n_i!} \frac{n_i!}{b_i^{n_i}\, e^{-b_i}} = \prod_i e^{-s_i} \left(1 + \frac{s_i}{b_i}\right)^{n_i}. \tag{A.1.4}$$

In fact the same information is contained in the distribution of the natural logarithm of this variable,

$$\log(q_W(\boldsymbol{n})) = \sum_i n_i \log\left(1 + \frac{s_i}{b_i}\right) - s_i. \tag{A.1.5}$$

For every choice of H_0 and H_1, the test statistic is a function of the "measurement", which may be a real observation or from a toy MC simulation. Therefore, this variable will have different probability mass function for the different hypotheses $P_0(q_W|H_0)$ and $P_1(q_W|H_1)$. In principle, since q_W is a function of $\boldsymbol{n}$ and we have the probability mass function over $\boldsymbol{n}$ for either of the two hypotheses, we can construct $P_0(q_W|H_0)$ and $P_1(q_W|H_1)$. Even though the values that the function can take are discrete, the discreteness does not present itself in a significant way, so from now on the probability mass function will be referred to as probability density over

continuous variables. With the two distributions at hand, for each experimental measurement we can evaluate the test statistic $q_W(\boldsymbol{n}_{\text{obs}}) \equiv \tilde{q}_W$ and find the type II error of falsely rejecting the alternative hypothesis

$$CL_{s+b} = P(q_W \leq \tilde{q}_W | H_1)\,. \tag{A.1.6}$$

One way to define the confidence level of rejecting the alternative hypothesis would be $(1 - CL_{s+b})$. This would be a reasonable value if the measurement falls within or beyond the bulk of the null hypothesis distribution. In a case of severe downward fluctuation, the exclusion $(1 - CL_{s+b})$ will be large for the alternative hypothesis, but the same would be true for the null hypothesis and $(1 - CL_{s+b})$ does not have any information about that. Therefore, to avoid concluding that an alternative hypothesis is false in an experiment that is an obvious outlier with respect to both H_0 and H_1, we use a definition of the confidence level that incorporates information relative to both distributions [118]

$$\begin{aligned} CL_s &= \frac{CL_{s+b}}{CL_b}\,, \text{ where} \\ CL_b &= P(q_W \leq \tilde{q}_W | H_0)\,. \end{aligned} \tag{A.1.7}$$

The confidence level that we quote is Sec. 3.2 is $(1 - CL_s) \times 100\%$. With this modification in the case of a downward background fluctuation, the experiment will not be able to quote a large confidence level for exclusion.

The explanation so far has an implicit assumption that the expected events under the alternative hypothesis are more than the null hypothesis mean. In our application this means $\Delta f > 0$. But the first result that we interpret is for $\Delta f = -1$, or the exclusion of the QCD-only hypothesis given the full SM as a null hypothesis. It is not difficult to show how the same interpretation can be achieved by noting that the log likelihood ratio is exactly the negative of what it would be if the expected difference between H_0 and H_a has the same magnitude but is positive. Essentially, the probability density of the null hypothesis will correspond to larger values of q_W than the alternative hypothesis. Therefore, the confidence level can be estimated if we change the direction of the inequality sign from $q_W \leq \tilde{q}_W$ to $q_W \geq \tilde{q}_W$. Of course in our analysis we do not have a real $\tilde{q}_W$. Rather we calculate the expected $(1 - CL_s)$ for q_W corresponding to the median of the H_0 distribution.

What is left is the technical issue of actually computing the probability densities of the test statistic under the two hypotheses. This is done by generating two sets of $\boldsymbol{n}_j$ toy "observations", one set following the statistical model in Eq. A.1.3 for the H_0 and the other for the H_1 hypothesis. For each "observation" j the test statistic is computed q_W^j. We order the two sets in ascending order, which allows us to easily calculate probabilities like $P(q_W \leq \tilde{q}_{\mathrm{W}}|H)$ by finding the nearest to $\tilde{q}_{\mathrm{W}}$ member of the set corresponding to hypothesis H, q_W^j, and dividing the ordered index j by the total number of toys.

So far the statistical picture that we have used is purely frequentist. For the inclusion of a systematic uncertainty on the background expectation value, we follow the prescription in [118], which is Bayesian in nature. The updated statistical model that describes the probability density to get a result n_i in bin i is

$$P(n_i|s_i, b_i) = \frac{\int db_i' P_c(n_i|s_i, b_i') P_s(b_i'|b_i)}{\int db_i' P_s(b_i'|b_i)}. \tag{A.1.8}$$

In Bayesian terminology, we define a probability distribution of the true parameter b_i' that depends on the value calculated from theory b_i. This probability distribution is the prior and the Poisson model is the likelihood, which is a function of the true parameter b_i'. Marginalising over this parameter, we get a function proportional to the probability density over the possible observations n_i. The test statistic is changed to accommodate the new statistical model

$$q_W(\boldsymbol{n}) = \prod_i \frac{\int db_i' P_c(n_i|s_i, b_i') P_s(b_i'|b_i)}{\int db_i' P_c(n_i|b_i') P_s(b_i'|b_i)}. \tag{A.1.9}$$

If we do the marginalisation over the alternative hypothesis in the numerator and the null hypothesis in the denominator, we get a likelihood ratio. This notion of the true parameter as a probability distribution that reflects the best belief, is quintessentially Bayesian. Once we have a definition of the test statistic for each possible measurement and the associated probability of that measurement, we can construct the probability distributions of the test statistic for each of the two hypotheses and use the same inferential techniques as we did for a known background. Practically, before each toy "observation" we first select two values according to the normal distribution $\mathcal{N}(1, \epsilon)$. Then we multiply all b_i by one of them to generate all b_i' for

$P_c(n_i|s_i, b'_i)$ and we do the same with the other number to get $P_c(n_i|b'_i)$. Finally, we extract a single "observation" from these Poisson distributions. The steps are repeated for the next "observation".

A.2 $t\bar{t}H$ identification

To analyse and interpret the results from our $t\bar{t}H$ search, we use broadly the same framework of the modified frequentist method from the previous section. We combine the same building blocks into the statistical model of the results, we define a test statistic and generate its distributions under null and alternative hypotheses. From them we apply the CL_s procedure to define the confidence level at which models can be reject by the data. The similarities end here and the details of the procedure are very distinct. We adopt the methodology agreed by the ATLAS and CMS collaborations for the statistical interpretation of experimental results [139] with small modifications to fit the nature of our search and huge simplifications in the statistical models.

First, we define a signal strength modifier μ, common to all selection channels, that is defined as a proportion of the expected Standard Model $t\bar{t}H$ signal. Then, in every channel i, the null hypothesis is the sum of expected $t\bar{t} + X$ events, including $t\bar{t}H$, as predicted by the SM (denoted b_i). The alternative hypotheses are defined as $\mu s_i + b_i$, where s_i is the SM expected contribution from $t\bar{t}H$ events in that bin. With this notation, the null hypothesis (the total expectation from the SM) corresponds to $\mu = 0$. Any deviation from it is measured in proportion to the SM $t\bar{t}H$ expectation. Our analysis allows for $\mu < 0$, which is not incorporated into the definition for the original Higgs search in [39] for obvious reasons.

Previously, the systematic uncertainty was interpreted as a probability distribution over the true background-only expectation value. For this analysis we reinterpret this degree of belief as a posterior $P(b'|b_i)$ from an auxiliary measurement that finds b_i. Therefore, it is proportional to the product of the likelihood $P(b_i|b'_i)$ of getting the outcome b_i given the true value b'_i and a hyper-prior $P(b'_i)$ that can

be chosen to be minimally biased

$$P(b_i'|b_i) \propto P(b_i|b_i') \cdot P(b_i')\,. \tag{A.2.10}$$

With the Gaussian choice for the likelihood, a uniform hyper-prior and by re-labelling for convenience the true parameter as b_i and the auxiliary measurement as b_i', we get the posterior-likelihood relation $P_s(b_i|b_i') = P_s(b_i'|b_i)$. This re-interpretation allows to naturally incorporate the background uncertainty in a frequentist way to the Poisson statistical model of a counting experiment. We can build distributions of test statistics by sampling the improved statistical model instead of integrating over the background parameter first.

Therefore the statistical model for the possible results is

$$P_{\text{tot}}(\boldsymbol{n}|\mu, \boldsymbol{b}) = \prod_i P_c(n_i|\mu s_i + b_i) \cdot P(b_i'|b_i)\,. \tag{A.2.11}$$

From it we can sample the distribution of $\boldsymbol{n}$ for different μ. Moreover, we can use it to define a test statistic for each hypothesis H_μ. We borrow the profile likelihood ratio from [139]

$$q_\mu = -2 \log \frac{P_{\text{tot}}(\boldsymbol{n}|\mu, \hat{\boldsymbol{b}}_\mu)}{P_{\text{tot}}(\boldsymbol{n}|\hat{\mu}, \hat{\boldsymbol{b}})}. \tag{A.2.12}$$

The parameter $\hat{\boldsymbol{b}}_\mu$ is the one that maximises $P_{\text{tot}}(\boldsymbol{n}|\mu, \boldsymbol{b})$ for the "measured" data $\boldsymbol{n}$ and the signal strength (μ) we try to exclude. The denominator is the maximised $P_{\text{tot}}(\boldsymbol{n})$ for the "measurement" $\boldsymbol{n}$ over the full range of μ and $\boldsymbol{b}$.

Before evaluating confidence levels, the probability densities of the test statistic need to be determined. Previously, it was relatively easy to generate toy "observations" and evaluate q_W for each, because the value of the nuisance parameters was selected before the Poisson distribution was sampled. Now the observation is required in order to find the best parameter $\hat{\boldsymbol{b}}_\mu$ to define the sampling distribution, from which we extract toy "observations". In order to break this paradox, we have to supply the experimental observation. The obvious choice is the expected value for the $\mu = 0$ hypothesis. Now we have $\hat{b}_\mu'$ for the "observation" $\boldsymbol{n} = \boldsymbol{b}'$ and we are ready to create the probability distributions $P(q_\mu|\mu, \hat{b}_\mu')$ and $P(q_\mu|0, \hat{b}_0')$. For each luminosity, we test 51 models within a luminosity-specific range $[-\mu_l, \mu_l]$ that are

equally spaced. We do not perform the optimisation and MC sampling, but use the RooStats [137] package in the ROOT [136] framework.

With these distributions it is easy to extract $CL_{s+b} = P(q_\mu \geq \tilde{q}_\mu|\mu, \hat{\boldsymbol{b}}'_\mu)$ and $CL_b = P(q_\mu \geq \tilde{q}_\mu|0, \hat{\boldsymbol{b}}'_0)$. This allows us to construct a function $CL_s(\mu, \tilde{q}_\mu)$. We find the two values of μ (one positive one negative) that give $CL_s = 0.025$ for two special choices of $\tilde{q}_\mu$. One of them is the median of the background distribution and the other is a fluctuation by $1\,\sigma$. They provide the range of μ around the SM that would not be able to be excluded at 95% CL if the experimental observation falls exactly at the median of the null hypothesis $P(q_\mu \geq \tilde{q}_\mu|0, \hat{\boldsymbol{b}}'_0) = 50\%$ (green band in Fig. 4.13 and 4.14) or within 1σ (yellow band).

Appendix B

HEPTopTagger

The HEPTopTagger [132], is an algorithm that determines if a (very wide) jet contains the decay products of a hadronically decaying top. There are two main stages in the process of tagging. The first is a grooming procedure by which soft and spurious radiation is removed from within the jet. The second is a kinematic constraint on the remaining hard structures within the fat jet, which is justified by the two expected mass scales in a hadronic top decay - the top mass and the W mass.

The tagger is applied to a jet, so the first step is to define the fat jet. The only constraint here is that the jet radius needs to be large in order to have a reasonable chance of containing the remnants of the three quarks from the top decay. In our $t\bar{t}H$ analysis, we define the fat jets to be Cambridge/Aachen with $R = 1.5$ and $p_T > 200$ GeV.

The next step is a mass drop condition at each clustering step in order to remove structures that individually do not add significantly to the mass of the jet. Starting from the final jet j, separate the last step of the clustering process into the parent pseudo jets j_1 and j_2. The convention is $m_{j_1} > m_{j_2}$. If the large-mass pseudo jet satisfies $m_{j_1} < 0.8\, m_j$, or in other words when there is a significant drop in the mass scale, both pseudo jets are kept. Otherwise the softer pseudo jet is discarded. The procedure is repeated to each remaining pseudo jet with mass $m_{j_i} > 30$ GeV. If the pseudo jet's mass is less than than, it is not declustered any further, but is kept as a hard structure for later use.

At the end of the mass drop procedure, all remaining hard structures are grouped

into all possible combinations of three. The grooming proceed for each combination separately. The constituents in the group are filtered by clustering them into small C/A jets with $R = \min(0.3, \Delta R_{jk}/2)$, where ΔR_{jk} is the distance in $\eta - \phi$ between the pair of substructures (j, k). The hardest (up to) five filtered subjets form the top candidate associated with the group of three substructures. The masses of all top candidates are calculated and only the candidate with the closest mass to m_t is selected for the actual kinematic tagging.

The objects that will be used to assert if the kinematic conditions are satisfied are exactly three exclusive subjets, reconstructed from the constituents of the top candidate. They should correspond to the b quark and the two light quarks from the W decay. The problem is we do not know which pair of subjets corresponds to the W boson. In any case, the top is identified with the sum of the three jet momenta; therefore we have the condition $m_t^2 = m_{123}^2 = (p_1+p_2+p_3)^2$. In the limit where each $p_i^2 \approx 0$, this turns into $m_t^2 = (p_1+p_2)^2 + (p_1+p_3)^2 + (p_3+p_2)^2 = m_{12}^2 + m_{13}^2 + m_{23}^2$. Therefore, we can think of the three masses as x, y, z coordinates, and the top mass condition as an equation of a sphere with radius m_t in the space of m_{13}, m_{12}, m_{23}. Since the masses are positive, it is actually an eighth of a sphere. There are two degrees of freedom after the top mass constraint. Lets take one to be m_{23}/m_{123} (the z component), and the other - the azimuthal angle between the x axis (the m_{13} values) and the projection of the (m_{13}, m_{12}, m_{23}) vector in the $x - y$ plane, $\phi = \text{atan}\frac{m_{12}}{m_{13}}$.

Assuming one of the pairs matches exactly m_W, there are three possible relations between m_{23}/m_{123} and $\text{atan}(m_{12}/m_{13}) \equiv \phi$. If the W pair is j_2, j_3 then $m_{23}/m_{123} = m_W/m_t$ is a constant over all values of the angle $\text{atan}(m_{12}/m_{13}) \in [0, \pi/2]$. If the W pair is j_1, j_2, then the relation between m_{23}/m_{123} and ϕ is not trivial,

$$\begin{aligned} m_{123}^2 &= m_{12}^2 + m_{13}^2 + m_{23}^2 \\ \frac{m_{23}^2}{m_{123}^2} &= 1 - \frac{m_{12}^2}{m_{123}^2} - \frac{m_{13}^2}{m_{123}^2} \\ 1 - \frac{m_{23}^2}{m_{123}^2} &= \frac{m_{12}^2}{m_{123}^2}\left(1 + \frac{m_{13}^2}{m_{12}^2}\right) \\ 1 - \frac{m_{23}^2}{m_{123}^2} &= \frac{m_W^2}{m_t^2}\left(1 + \cot^2\phi\right). \end{aligned} \tag{B.0.1}$$

The last case $m_W = m_{13}$ is almost equivalent to the one above. Swapping m_{12} and m_{13} in the third line yields

$$1 - \frac{m_{23}^2}{m_{123}^2} = \frac{m_W^2}{m_t^2}\left(1 + \tan^2\phi\right). \tag{B.0.2}$$

Since the configuration of the top decay must fall into one of these three categories, the tagger can require that the three subjets satisfy any of three constraints:

$$\begin{aligned}
&1.\ R_{\text{min}} < \frac{m_{23}}{m_{123}} < R_{\text{max}} \ \text{ and } \ \phi \in [0.2, 1.3]\ ; \\
&2.\ R_{\text{min}}^2\left(1 + \cot^2\phi\right) < 1 - \frac{m_{23}^2}{m_{123}^2} < R_{\text{max}}^2\left(1 + \cot^2\phi\right) \ \text{ and } \ \frac{m_{23}}{m_{123}} > 0.35\,; \\
&2.\ R_{\text{min}}^2\left(1 + \tan^2\phi\right) < 1 - \frac{m_{23}^2}{m_{123}^2} < R_{\text{max}}^2\left(1 + \tan^2\phi\right) \ \text{ and } \ \frac{m_{23}}{m_{123}} > 0.35\,.
\end{aligned} \tag{B.0.3}$$

The ratios $R_{\text{min}} = 0.85\frac{m_W}{m_t}$ and $R_{\text{max}} = 1.15\frac{m_W}{m_t}$ define the band around m_W/m_t that the tagger considers acceptable. The ϕ range in the first line is constrained because neither m_{12} nor m_{13} can be zero, but are bound at around 30 GeV. The tagger also excludes regions where m_{23} is small.

Appendix C

Ellipticity

The ellipticity $\hat{t}$ of a jet is calculated from its particles' three-momentum components $\mathbf{k}_{Ti}$ transverse to the jet. Thus, it is defined in the plane transverse to the momentum $\mathbf{p}_J = \sum_i \mathbf{p}_i$, where $\mathbf{p}_i$ are the three-momenta of the jet constituents, as

$$\mathbf{k}_{Ti} \;=\; \mathbf{p}_i - (\mathbf{p}_J \cdot \mathbf{p}_i)\,\frac{\mathbf{p}_J}{|\mathbf{p}_J|^2}\;. \tag{C.0.1}$$

While we take $\mathbf{p}_J$ to be the thrust axis, we calculate thrust major T_{maj} and thrust minor T_{min} using the $\mathbf{k}_{Ti}$ as input

$$T_{\mathrm{maj}} \;=\; \max_{\mathbf{n}_{\mathrm{maj}}} \frac{\sum_i |\mathbf{k}_{Ti}\cdot\mathbf{n}_{\mathrm{maj}}|}{\sum_i |\mathbf{p}_{Ti}|} \qquad \text{and} \qquad T_{\mathrm{min}} \;=\; \frac{\sum_i |\mathbf{k}_{Ti}\cdot\mathbf{n}_{\mathrm{min}}|}{\sum_i |\mathbf{p}_{Ti}|}\;, \tag{C.0.2}$$

where $\mathbf{n}^2_{\mathrm{maj}} = \mathbf{n}^2_{\mathrm{min}} = 1$, $\mathbf{n}_{\mathrm{min}}\cdot\mathbf{n}_{\mathrm{maj}} = 0$ and $\mathbf{n}_{\mathrm{min}}\cdot\mathbf{p}_J = 0$. We then define the ellipticity as the ratio

$$\hat{t} \;=\; \frac{T_{\mathrm{min}}}{T_{\mathrm{maj}}}\;. \tag{C.0.3}$$

The two limiting cases are homogeneously distributed radiation within the jet cone (a circle in the transverse plane) and a planar distribution of the radiation within the cone (a line in the transverse plane). The former gives $\hat{t} = 1$ and the latter - $\hat{t} = 0$.

Acknowledgements

I would like to thank the Science and Technology Facilities Council (STFC) for funding my postgraduate studies and research. In addition, the completion of this thesis would have been impossible without the help and support of many people.

First and foremost, I am very grateful to my supervisor, Michael Spannowsky, who was always ready with advice and guidance when my progress would halt. Not to mention that, no matter the task at hand, he would always dig up a useful snippet of code from the past.

I am also thankful to all my other collaborators - Davison Soper, Danilo Ferreira de Lima, Stefano Pozzorini, Niccolo Moretti, Marek Schonherr and Frank Krauss - for taking the time to teach me different concepts, relevant to our research, and helping out when my technical skills needed assistance. Special thanks to Frank, who not only taught me a great deal of physics, but encouraged me to pursue a Ph.D. in particle physics when I was an undergraduate.

Some of the things I will always remember fondly are the conversations during lunchtime at the IPPP, which would very quickly wonder off from reasonable science into a mesh of topics that only made sense in that time and place. Many thanks to my postgraduate colleagues for those and other enjoyable conversations.

Finally, I cannot overstate the support from my family and girlfriend, who were always certain I would be able to carry through the research to the end and reminded me just often enough.

Bibliography

[1] D. Ferreira de Lima, P. Petrov, D. Soper, and M. Spannowsky, "Quark-Gluon tagging with Shower Deconstruction: Unearthing dark matter and Higgs couplings," 2016.

[2] F. Krauss, P. Petrov, M. Schoenherr, and M. Spannowsky, "Measuring collinear W emissions inside jets," *Phys. Rev.*, vol. D89, no. 11, p. 114006, 2014.

[3] N. Moretti, P. Petrov, S. Pozzorini, and M. Spannowsky, "Measuring the signal strength in $t\bar{t}H$ with $H \to b\bar{b}$," *Phys. Rev.*, vol. D93, no. 1, p. 014019, 2016.

[4] M. Lepka, "Cms slice," 2010.

[5] D. E. Soper and M. Spannowsky, "Finding physics signals with shower deconstruction," *Phys. Rev.*, vol. D84, p. 074002, 2011.

[6] S. L. Glashow, "Partial Symmetries of Weak Interactions," *Nucl. Phys.*, vol. 22, pp. 579–588, 1961.

[7] S. Weinberg, "A Model of Leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, 1967.

[8] A. Salam, "Weak and Electromagnetic Interactions," *Conf. Proc.*, vol. C680519, pp. 367–377, 1968.

[9] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, "Advantages of the Color Octet Gluon Picture," *Phys. Lett.*, vol. B47, pp. 365–368, 1973.

[10] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," *Phys. Rev. Lett.*, vol. 13, pp. 321–323, 1964.

[11] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 1964.

[12] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles," *Phys. Rev. Lett.*, vol. 13, pp. 585–587, 1964.

[13] G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett.*, vol. B716, pp. 1–29, 2012.

[14] S. Chatrchyan *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett.*, vol. B716, pp. 30–61, 2012.

[15] G. Bertone, D. Hooper, and J. Silk, "Particle dark matter: Evidence, candidates and constraints," *Phys. Rept.*, vol. 405, pp. 279–390, 2005.

[16] Z. L. A. Ceccuc and Y. Sakai, "The ckm quark-mixing matrix," *PDG Review*, 2014.

[17] Q. R. Ahmad *et al.*, "Measurement of the rate of $\nu_e + d \rightarrow p + p + e^-$ interactions produced by ^8B solar neutrinos at the Sudbury Neutrino Observatory," *Phys. Rev. Lett.*, vol. 87, p. 071301, 2001.

[18] D. E. Soper and M. Spannowsky, "Finding top quarks with shower deconstruction," *Phys. Rev.*, vol. D87, p. 054012, 2013.

[19] ——, "Finding physics signals with event deconstruction," *Phys. Rev.*, vol. D89, no. 9, p. 094005, 2014.

[20] M. Gell-Mann, "A Schematic Model of Baryons and Mesons," *Phys. Lett.*, vol. 8, pp. 214–215, 1964.

[21] G. Zweig, "An SU(3) model for strong interaction symmetry and its breaking. Version 2," in *DEVELOPMENTS IN THE QUARK*

THEORY OF HADRONS. VOL. 1. 1964 - 1978, D. Lichtenberg and S. P. Rosen, Eds., 1964, pp. 22–101. [Online]. Available: http://inspirehep.net/record/4674/files/cern-th-412.pdf

[22] M. Y. Han and Y. Nambu, "Three Triplet Model with Double SU(3) Symmetry," *Phys. Rev.*, vol. 139, pp. B1006–B1010, 1965.

[23] O. W. Greenberg, "Spin and Unitary Spin Independence in a Paraquark Model of Baryons and Mesons," *Phys. Rev. Lett.*, vol. 13, pp. 598–602, 1964.

[24] R. K. Ellis, W. J. Stirling, and B. R. Webber, "QCD and collider physics," *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.*, vol. 8, pp. 1–435, 1996.

[25] V. V. Ezhela, S. B. Lugovsky, and O. V. Zenin, "Hadronic part of the muon g-2 estimated on the sigma**2003(tot)(e+ e- —¿ hadrons) evaluated data compilation," 2003.

[26] E. D. Bloom *et al.*, "High-Energy Inelastic e p Scattering at 6-Degrees and 10-Degrees," *Phys. Rev. Lett.*, vol. 23, pp. 930–934, 1969.

[27] J. D. Bjorken and E. A. Paschos, "Inelastic Electron Proton and gamma Proton Scattering, and the Structure of the Nucleon," *Phys. Rev.*, vol. 185, pp. 1975–1982, 1969.

[28] C. G. Callan, Jr. and D. J. Gross, "High-energy electroproduction and the constitution of the electric current," *Phys. Rev. Lett.*, vol. 22, pp. 156–159, 1969.

[29] C. H. Llewellyn Smith, "INELASTIC LEPTON SCATTERING IN GLUON MODELS," *Phys. Rev.*, vol. D4, p. 2392, 1971.

[30] K. G. Wilson, "Confinement of Quarks," *Phys. Rev.*, vol. D10, pp. 2445–2459, 1974, [,45(1974)].

[31] S. Durr *et al.*, "Ab-Initio Determination of Light Hadron Masses," *Science*, vol. 322, pp. 1224–1227, 2008.

[32] H. J. Rothe, "Lattice gauge theories: An Introduction," *World Sci. Lect. Notes Phys.*, vol. 43, pp. 1–381, 1992, [World Sci. Lect. Notes Phys.82,1(2012)].

[33] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*, 1995. [Online]. Available: http://www.slac.stanford.edu/spires/find/books/www?cl=QC174.45%3AP4

[34] L. D. Faddeev and V. N. Popov, "Feynman Diagrams for the Yang-Mills Field," *Phys. Lett.*, vol. B25, pp. 29–30, 1967.

[35] C. G. Callan, Jr., "Broken scale invariance in scalar field theory," *Phys. Rev.*, vol. D2, pp. 1541–1547, 1970.

[36] K. Symanzik, "Small distance behavior in field theory and power counting," *Commun. Math. Phys.*, vol. 18, pp. 227–246, 1970.

[37] S. Chatrchyan *et al.*, "The CMS experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.

[38] G. Aad *et al.*, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08003, 2008.

[39] V. Khachatryan *et al.*, "Observation of the diphoton decay of the Higgs boson and measurement of its properties," *Eur. Phys. J.*, vol. C74, no. 10, p. 3076, 2014.

[40] G. Aad *et al.*, "Improved luminosity determination in pp collisions at sqrt(s) = 7 TeV using the ATLAS detector at the LHC," *Eur. Phys. J.*, vol. C73, no. 8, p. 2518, 2013.

[41] H. Lehmann, K. Symanzik, and W. Zimmermann, "On the formulation of quantized field theories," *Nuovo Cim.*, vol. 1, pp. 205–225, 1955.

[42] F. Bloch and A. Nordsieck, "Note on the Radiation Field of the electron," *Phys. Rev.*, vol. 52, pp. 54–59, 1937.

[43] T. Kinoshita, "Mass singularities of Feynman amplitudes," *J. Math. Phys.*, vol. 3, pp. 650–677, 1962.

[44] T. D. Lee and M. Nauenberg, "Degenerate Systems and Mass Singularities," *Phys. Rev.*, vol. 133, pp. B1549–B1562, 1964, [,25(1964)].

[45] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, "New clustering algorithm for multi - jet cross-sections in e+ e- annihilation," *Phys. Lett.*, vol. B269, pp. 432–438, 1991.

[46] M. Cacciari, G. P. Salam, and G. Soyez, "The Anti-k(t) jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008.

[47] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms," *JHEP*, vol. 08, p. 001, 1997.

[48] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language," *Nucl. Phys.*, vol. B126, pp. 298–318, 1977.

[49] R. P. Feynman, "Very high-energy collisions of hadrons," *Phys. Rev. Lett.*, vol. 23, pp. 1415–1417, 1969.

[50] M. Breidenbach, J. I. Friedman, H. W. Kendall, E. D. Bloom, D. H. Coward, H. C. DeStaebler, J. Drees, L. W. Mo, and R. E. Taylor, "Observed Behavior of Highly Inelastic electron-Proton Scattering," *Phys. Rev. Lett.*, vol. 23, pp. 935–939, 1969.

[51] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." *Sov. Phys. JETP*, vol. 46, pp. 641–653, 1977, [Zh. Eksp. Teor. Fiz.73,1216(1977)].

[52] V. N. Gribov and L. N. Lipatov, "Deep inelastic e p scattering in perturbation theory," *Sov. J. Nucl. Phys.*, vol. 15, pp. 438–450, 1972, [Yad. Fiz.15,781(1972)].

[53] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual," *JHEP*, vol. 05, p. 026, 2006.

[54] M. Bahr *et al.*, "Herwig++ Physics and Manual," *Eur. Phys. J.*, vol. C58, pp. 639–707, 2008.

[55] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, "Event generation with SHERPA 1.1," *JHEP*, vol. 02, p. 007, 2009.

[56] A. H. Mueller, "Multiplicity and Hadron Distributions in QCD Jets: Nonleading Terms," *Nucl. Phys.*, vol. B213, pp. 85–108, 1983.

[57] H. Georgi and M. Machacek, "A Simple QCD Prediction of Jet Structure in e+ e- Annihilation," *Phys. Rev. Lett.*, vol. 39, p. 1237, 1977.

[58] S. Catani, G. Turnock, and B. R. Webber, "Jet broadening measures in e^+e^- annihilation," *Phys. Lett.*, vol. B295, pp. 269–276, 1992.

[59] J. Thaler and K. Van Tilburg, "Identifying Boosted Objects with N-subjettiness," *JHEP*, vol. 03, p. 015, 2011.

[60] J. Gallicchio and M. D. Schwartz, "Quark and Gluon Tagging at the LHC," *Phys. Rev. Lett.*, vol. 107, p. 172001, 2011.

[61] A. J. Larkoski, G. P. Salam, and J. Thaler, "Energy Correlation Functions for Jet Substructure," *JHEP*, vol. 06, p. 108, 2013.

[62] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, "Gaining (Mutual) Information about Quark/Gluon Discrimination," *JHEP*, vol. 11, p. 129, 2014.

[63] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, "Associated jet and subjet rates in light-quark and gluon jet discrimination," *JHEP*, vol. 04, p. 131, 2015.

[64] J. R. Andersen *et al.*, "Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report," in *9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, June 1-19, 2015*, 2016. [Online]. Available: http://lss.fnal.gov/archive/2016/conf/fermilab-conf-16-175-ppd-t.pdf

[65] G. Aad *et al.*, "Light-quark and gluon jet discrimination in *pp* collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Eur. Phys. J.*, vol. C74, no. 8, p. 3023, 2014.

[66] "Performance of quark/gluon discrimination in 8 TeV pp data," CERN, Geneva, Tech. Rep. CMS-PAS-JME-13-002, 2013. [Online]. Available: https://cds.cern.ch/record/1599732

[67] V. Khachatryan *et al.*, "Search for dark matter, extra dimensions, and unparticles in monojet events in protonproton collisions at $\sqrt{s} = 8$ TeV," *Eur. Phys. J.*, vol. C75, no. 5, p. 235, 2015.

[68] Y. L. Dokshitzer, V. A. Khoze, and T. Sjostrand, "Rapidity gaps in Higgs production," *Phys. Lett.*, vol. B274, pp. 116–121, 1992.

[69] D. L. Rainwater, D. Zeppenfeld, and K. Hagiwara, "Searching for $H \to \tau^+\tau^-$ in weak boson fusion at the CERN LHC," *Phys. Rev.*, vol. D59, p. 014037, 1998.

[70] D. Zeppenfeld, R. Kinnunen, A. Nikitenko, and E. Richter-Was, "Measuring Higgs boson couplings at the CERN LHC," *Phys. Rev.*, vol. D62, p. 013009, 2000.

[71] C. Englert, R. Kogler, H. Schulz, and M. Spannowsky, "Higgs coupling measurements at the LHC," 2015.

[72] V. Del Duca, W. Kilgore, C. Oleari, C. Schmidt, and D. Zeppenfeld, "Gluon fusion contributions to H + 2 jet production," *Nucl. Phys.*, vol. B616, pp. 367–399, 2001.

[73] J. Neyman and E. S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*. New York, NY: Springer New York, 1992, pp. 73–108. [Online]. Available: http://dx.doi.org/10.1007/978-1-4612-0919-5_6

[74] F. Fiedler, A. Grohsjean, P. Haefner, and P. Schieferdecker, "The Matrix Element Method and its Application in Measurements of the Top Quark Mass," *Nucl. Instrum. Meth.*, vol. A624, pp. 203–218, 2010.

[75] J. Alwall, A. Freitas, and O. Mattelaer, "The Matrix Element Method and QCD Radiation," *Phys. Rev.*, vol. D83, p. 074010, 2011.

[76] Z. Nagy and D. E. Soper, "Parton showers with quantum interference," *JHEP*, vol. 09, p. 114, 2007.

[77] ——, "Parton showers with quantum interference: Leading color, spin averaged," *JHEP*, vol. 03, p. 030, 2008.

[78] ——, "Parton showers with quantum interference: Leading color, with spin," *JHEP*, vol. 07, p. 025, 2008.

[79] S. Ask, J. H. Collins, J. R. Forshaw, K. Joshi, and A. D. Pilkington, "Identifying the colour of TeV-scale resonances," *JHEP*, vol. 01, p. 018, 2012.

[80] K. Joshi, A. D. Pilkington, and M. Spannowsky, "The dependency of boosted tagging algorithms on the event colour structure," *Phys. Rev.*, vol. D86, p. 114016, 2012.

[81] G. Aad *et al.*, "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1," 2016.

[82] "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET," CERN, 2009. Geneva, Tech. Rep. CMS-PAS-PFT-09-001, Apr 2009. [Online]. Available: https://cds.cern.ch/record/1194487

[83] "Jet Energy Corrections for Multiple Cone Sizes," CERN, 2009. Geneva, Tech. Rep. [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/MultipleConeSizes14

[84] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 1992.

[85] A. J. Larkoski, I. Moult, and D. Neill, "Power Counting to Better Jet Observables," *JHEP*, vol. 12, p. 009, 2014.

[86] M. R. Buckley, D. Feld, and D. Goncalves, "Scalar Simplified Models for Dark Matter," *Phys. Rev.*, vol. D91, p. 015017, 2015.

[87] P. Harris, V. V. Khoze, M. Spannowsky, and C. Williams, "Constraining Dark Sectors at Colliders: Beyond the Effective Theory Approach," *Phys. Rev.*, vol. D91, p. 055009, 2015.

[88] M. Aaboud *et al.*, "Search for new phenomena in final states with an energetic jet and large missing transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector," 2016.

[89] D. L. Rainwater and D. Zeppenfeld, "Observing $H \to W^*W^* \to e^{\pm}\mu\mp \not{p}_T$ in weak boson fusion with dual forward jet tagging at the CERN LHC," *Phys. Rev.*, vol. D60, p. 113004, 1999, [Erratum: Phys. Rev.D61,099901(2000)].

[90] B. E. Cox, J. R. Forshaw, and A. D. Pilkington, "Extracting Higgs boson couplings using a jet veto," *Phys. Lett.*, vol. B696, pp. 87–91, 2011.

[91] J. R. Andersen, C. Englert, and M. Spannowsky, "Extracting precise Higgs couplings by using the matrix element method," *Phys. Rev.*, vol. D87, no. 1, p. 015019, 2013.

[92] C. Englert, M. Spannowsky, and M. Takeuchi, "Measuring Higgs CP and couplings with hadronic event shapes," *JHEP*, vol. 06, p. 108, 2012.

[93] P. Ciafaloni and D. Comelli, "Sudakov enhancement of electroweak corrections," *Phys.Lett.*, vol. B446, pp. 278–284, 1999.

[94] J. H. Kuhn, A. Penin, and V. A. Smirnov, "Summing up subleading Sudakov logarithms," *Eur.Phys.J.*, vol. C17, pp. 97–105, 2000.

[95] V. S. Fadin, L. Lipatov, A. D. Martin, and M. Melles, "Resummation of double logarithms in electroweak high-energy processes," *Phys.Rev.*, vol. D61, p. 094002, 2000.

[96] M. Ciafaloni, P. Ciafaloni, and D. Comelli, "Bloch-Nordsieck violating electroweak corrections to inclusive TeV scale hard processes," *Phys.Rev.Lett.*, vol. 84, pp. 4810–4813, 2000.

[97] M. Hori, H. Kawamura, and J. Kodaira, "Electroweak Sudakov at two loop level," *Phys.Lett.*, vol. B491, pp. 275–279, 2000.

[98] A. Denner and S. Pozzorini, "One loop leading logarithms in electroweak radiative corrections. 1. Results," *Eur.Phys.J.*, vol. C18, pp. 461–480, 2001.

[99] E. Accomando, A. Denner, and S. Pozzorini, "Electroweak correction effects in gauge boson pair production at the CERN LHC," *Phys.Rev.*, vol. D65, p. 073003, 2002.

[100] W. Beenakker and A. Werthenbach, "Electroweak two loop Sudakov logarithms for on-shell fermions and bosons," *Nucl.Phys.*, vol. B630, pp. 3–54, 2002.

[101] A. Denner, M. Melles, and S. Pozzorini, "Two loop electroweak angular dependent logarithms at high-energies," *Nucl.Phys.*, vol. B662, pp. 299–333, 2003.

[102] A. Denner and S. Pozzorini, "An Algorithm for the high-energy expansion of multi-loop diagrams to next-to-leading logarithmic accuracy," *Nucl.Phys.*, vol. B717, pp. 48–85, 2005.

[103] A. Denner, B. Jantzen, and S. Pozzorini, "Two-loop electroweak next-to-leading logarithms for processes involving heavy quarks," *JHEP*, vol. 0811, p. 062, 2008.

[104] U. Baur, "Weak Boson Emission in Hadron Collider Processes," *Phys.Rev.*, vol. D75, p. 013005, 2007.

[105] G. Bell, J. Kuhn, and J. Rittinger, "Electroweak Sudakov Logarithms and Real Gauge-Boson Radiation in the TeV Region," *Eur.Phys.J.*, vol. C70, pp. 659–671, 2010.

[106] J. R. Christiansen and T. Sjostrand, "Weak Gauge Boson Radiation in Parton Showers," 2014.

[107] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1," *Comput.Phys.Commun.*, vol. 178, pp. 852–867, 2008.

[108] T. Gleisberg and S. Höche, "Comix, a new matrix element generator," *JHEP*, vol. 0812, p. 039, 2008.

[109] Z. Nagy and D. E. Soper, "Matching parton showers to NLO computations," *JHEP*, vol. 0510, p. 024, 2005.

[110] S. Schumann and F. Krauss, "A Parton shower algorithm based on Catani-Seymour dipole factorisation," *JHEP*, vol. 0803, p. 038, 2008.

[111] J.-C. Winter, F. Krauss, and G. Soff, "A Modified cluster hadronization model," *Eur.Phys.J.*, vol. C36, pp. 381–395, 2004.

[112] S. Alekhin, G. Altarelli, N. Amapane, J. Andersen, V. Andreev *et al.*, "HERA and the LHC: A Workshop on the implications of HERA for LHC physics: Proceedings Part A," 2005.

[113] A. Denner and F. Hebenstreit, *unpublished*, 2006.

[114] S. Catani and M. Seymour, "A General algorithm for calculating jet cross-sections in NLO QCD," *Nucl.Phys.*, vol. B485, pp. 291–419, 1997.

[115] S. Catani, S. Dittmaier, M. H. Seymour, and Z. Trocsanyi, "The Dipole formalism for next-to-leading order QCD calculations with massive partons," *Nucl.Phys.*, vol. B627, pp. 189–265, 2002.

[116] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC," *Phys.Rev.Lett.*, vol. 100, p. 242001, 2008.

[117] K. Rehermann and B. Tweedie, "Efficient Identification of Boosted Semileptonic Top Quarks at the LHC," *JHEP*, vol. 1103, p. 059, 2011.

[118] T. Junk, "Confidence level computation for combining searches with small statistics," *Nucl.Instrum.Meth.*, vol. A434, pp. 435–443, 1999.

[119] G. Aad *et al.*, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data," *Phys. Lett.*, vol. B726, pp. 120–144, 2013.

[120] T. ATLAS and C. Collaborations, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV," 2015.

[121] R. Lafaye, T. Plehn, M. Rauch, D. Zerwas, and M. Duhrssen, "Measuring the Higgs Sector," *JHEP*, vol. 08, p. 009, 2009.

[122] D. E. Soper and M. Spannowsky, "Combining subjet algorithms to enhance ZH detection at the LHC," *JHEP*, vol. 08, p. 029, 2010.

[123] T. Plehn, G. P. Salam, and M. Spannowsky, "Fat Jets for a Light Higgs," *Phys. Rev. Lett.*, vol. 104, p. 111801, 2010.

[124] G. Degrassi, S. Di Vita, J. Elias-Miro, J. R. Espinosa, G. F. Giudice, G. Isidori, and A. Strumia, "Higgs mass and vacuum stability in the Standard Model at NNLO," *JHEP*, vol. 08, p. 098, 2012.

[125] D. Buttazzo, G. Degrassi, P. P. Giardino, G. F. Giudice, F. Sala, A. Salvio, and A. Strumia, "Investigating the near-criticality of the Higgs boson," *JHEP*, vol. 12, p. 089, 2013.

[126] M. J. Dolan, C. Englert, and M. Spannowsky, "Higgs self-coupling measurements at the LHC," *JHEP*, vol. 10, p. 112, 2012.

[127] V. Khachatryan *et al.*, "Search for a Standard Model Higgs Boson Produced in Association with a Top-Quark Pair and Decaying to Bottom Quarks Using a Matrix Element Method," *Eur. Phys. J.*, vol. C75, no. 6, p. 251, 2015.

[128] G. Aad *et al.*, "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *Eur. Phys. J.*, vol. C75, no. 7, p. 349, 2015.

[129] V. Khachatryan *et al.*, "Search for the associated production of the Higgs boson with a top-quark pair," *JHEP*, vol. 09, p. 087, 2014, [Erratum: JHEP10,106(2014)].

[130] G. Aad *et al.*, "Search for the associated production of the Higgs boson with a top quark pair in multilepton final states with the ATLAS detector," *Phys. Lett.*, vol. B749, pp. 519–541, 2015.

[131] ——, "Search for $H \to \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector," *Phys. Lett.*, vol. B740, pp. 222–242, 2015.

[132] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, "Stop Reconstruction with Tagged Tops," *JHEP*, vol. 10, p. 078, 2010.

[133] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, "Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks," *Phys. Rev. Lett.*, vol. 101, p. 142001, 2008.

[134] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002200009791504X

[135] P. Speckmayer, A. Hocker, J. Stelzer, and H. Voss, "The toolkit for multivariate data analysis, TMVA 4," *J. Phys. Conf. Ser.*, vol. 219, p. 032057, 2010.

[136] I. Antcheva *et al.*, "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization," *Comput. Phys. Commun.*, vol. 182, pp. 1384–1385, 2011.

[137] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, G. Schott, W. Verkerke, and M. Wolf, "The RooStats Project," *PoS*, vol. ACAT2010, p. 057, 2010.

[138] S. Chatrchyan *et al.*, "Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV," *Phys. Lett.*, vol. B710, pp. 26–48, 2012.

[139] "Procedure for the LHC Higgs boson search combination in summer 2011," 2011.